A Tensor Encoding Model of Word Meaning

Theory and Application to Information Retrieval

A THESIS SUBMITTED TO

THE SCIENCE AND ENGINEERING FACULTY

OF QUEENSLAND UNIVERSITY OF TECHNOLOGY

IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Michael Symonds

Science and Engineering Faculty

Queensland University of Technology

May 2013

Copyright in Relation to This Thesis

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature:

To Jacquie and my parents

Abstract

The ability to create meaning from our environment is fundamental to the way in which we interact with our world. Research into human cognition and language has inspired the development of a class of computational models that induce meaning based on how words are associated in streams of natural language. These models represent semantic associations between words within a geometric space, hence their popular name, semantic space models. Their low cost and ability to provide effective performance on a number of linguistic tasks, including synonym judgement and analogy making, has seen them promoted as having potentially transformative powers for improving communication between humans and computers.

This thesis develops a formal model of word meaning, that extends recent advances in semantic space technology, including the use of tensor representations, and evaluates its performance on a number of semantic tasks, including synonym judgement, semantic distance, semantic categorization and similarity judgement of medical concepts. The model's formal framework, combined with its superior efficiency and effectiveness compared to existing models, provides a theoretical platform from which it is applied to the query expansion process within the information retrieval problem. Its effectiveness in this role is evaluated on several ad hoc retrieval and web search tasks. The demonstrated significant improvements in retrieval effectiveness of this approach over benchmark models motivates a discussion of other potential applications and areas for future work.

Acknowledgments

I would like to thank the many people who have provided valuable support and guidance during my Ph.D, especially Peter Bruza, my principal supervisor. His passion for research and surfing is always a source of inspiration. However, after several years of surf supervisor meetings I can safely say my research skills have improved significantly more than my surfing skills. Thank you Peter.

I would also like to thank my associate supervisors: Laurianne Sitbon, Ian Turner and my pseudo supervisor: Guido Zuccon, for their many discussions, reviews and supportive words along the way. Many thanks to Bevan Koopman and Anthony Nguyen for their contributions and reviews on a number of papers. I am grateful to Sándor Darányi for his review comments of this thesis. A special thanks to Peter Kaye for his editing skills on this dissertation and other papers, along with his many helpful insights.

To the members of the Quantum Interaction (QI) group at QUT, thank you all for the many thought provoking discussions and good times[1]. I would also like to thank the researchers who contributed data sets via papers or directly, including Bevan Koopman, Guido Zuccon, John Bullinaria and Tom Landauer.

Finally, I would like to thank my family and friends for their support on this journey. I am most grateful to my parents, Peggy and Jeff for their unconditional support and to my partner Jacquie for her love and understanding.

[1] http://www.youtube.com/watch?v=jrw40BJRzrs

Table of Contents

Nomenclature

Abbreviations

BEAGLE	Bound Encoding of the Aggregate Language Environment
BNC	British National Corpus
CF	Co-occurrence Frequency
CFC	Co-occurrence Frequency Cut-off
COALS	Correlated Occurrence Analogue to Lexical Semantic
DMM	Distributional Memory Model
fMRI	functional Magnetic Resonance Imaging
HAL	Hyperspace Analogue to Language
HRR	Holographic Reduced Representation
LCE	Latent Concept Expansion
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MRF	Markov Random Field
NLP	Natural Language Processing
POS	Part of Speech
PMI	Pointwise Mutual Information
PPMI	Positive Pointwise Mutual Information
PRI	Permuted Random Indexing
PRM	Positional Relevance Model
RI	Random Indexing

RM3	Unigram relevance model (using Dirichlet smoothing)
SSM	Semantic Space Model
SVD	Singular Value Decomposition
TASA	Touchstone Applied Science Associates (Corpus)
TE	Tensor Encoding (model)
TMC	Tensor Memory Compression
TOEFL	Test Of English as a Foreign Language
TQE	Tensor Query Expansion
TREC	Text REtrieval Conference
UMLS	Unified Medical Language System

Symbols		**Chapter**
$\otimes$	Outer product of vectors, or Kronecker product of tensors	Ch2,3
$\odot$	Circular convolution	Ch2

Greek Letters

α	Mix of original query weights in the relevance model	Ch5
γ	Mix of syntagmatic and paradigmatic information	Ch3,4,6,7,8
λ	Length of the BEAGLE context window used in binding	Ch2
λ	Jelinick-Mercer smoothing parameter	Ch5
$\vec{\Phi}$	Constant random placeholder vector used in BEAGLE	Ch2
σ	Standard deviation of a Gaussian distribution	Ch2
μ	Mean of a Gaussian distribution	Ch2
μ	Dirichlet smoothing parameter	Ch5,6,7

Subscripts

SV	Storage vector	Ch3,4,6,7

$para$	Paradigmatic	Ch3,4,6,7
syn	Syntagmatic	Ch3,4,6,7

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Word Meaning

The ability to create meaning from our environment is fundamental to the way in which we interact with our world. How information about the environment is represented within human cognition, and how meaning is then created has been an area of focused research for many decades.

So what is *meaning*? From a philosophical point of view, there is unlikely to ever be one agreed answer. From a practical point of view, relating to development of the work presented in this thesis, a pragmatic answer that explains how words take on meaning is required. To this end, the theories of structural linguistics provide a distributional view of meaning that is partially captured in J. R. Firth's (1957) popularised quote:

> *A word shall be known by the company it keeps.*

This infers that words that occur in similar contexts tend to have similar meanings. This inference is known as the *distributional hypothesis*, as the meaning of a word can be derived from the distribution of words surrounding it in natural language [Firth, 1957, Harris, 1954]. Structural linguistics goes further to explicitly link the meaning of a word not only to its relationship with neighbouring words, known as syntagmatic associations, but also the relationships formed between words that have common neighbours, known as paradigmatic associations.

More formally, a *syntagmatic association* exists between two words if they co-occur in natural language more frequently than expected from chance [Lyons, 1968]. Typically, syntagmatic relations exist between words from different parts of speech, i.e., *adjective-noun*. However,

words from the same part of speech can also display syntactic associations, e.g., "jungle-tiger" and "desert-oasis".

A *paradigmatic association* exists between two words if they can substitute for one another in a sentence without affecting the acceptability of the sentence [Lyons, 1968]. Typically, paradigmatic associations exist between words from the same part of speech [Rapp, 2002], such as related nouns (i.e., "tiger-panther") or related verbs (i.e., "slept-ran").

The idea that the meaning of a word can be induced from its syntagmatic and paradigmatic associations was first developed by Swiss linguist, Ferdinand de Saussure (1916). The ideas underpinning this "differential view of meaning" [Elffers, 2008, Pavel, 2001, Sahlgren, 2006] have been adapted by a number of prominent linguists, including work on *sense relations* by Lyons [1968], and have been argued to form a relatively clean theory of linguistics, free of psychology, sociology and anthropology [Holland, 1992]. Therefore, it is argued that structural linguistics is a well accepted theory which provides a relatively unobstructed path toward developing computational models of word meaning.

Given this theoretical setting, the ability to model *word* meaning becomes heavily dependent on identifying statistical relationships between words. Modelling these relationships based on the co-occurrence patterns of words in natural language has been achieved within geometric [Sahlgren, 2006, Schütze and Pedersen, 1993] and probabilistic [Weeds, 2003, Yuret, 1998] settings. The intimate relationship between geometry and meaning has been recently articulated by a number of prominent researchers, including van Rijsbergen [2004] and Widdows [2004].

1.2 The Geometry of Meaning

A successful approach to modelling meaning involves representing relationships between words in a geometric space. An early example, which has made a valuable contribution to the field of psychology, is the work on the *semantic differential* by Osgood et al. [1957]. Osgood modelled the affective (emotional) meaning of words based on a number of bipolar scales; each scale acting as a dimension in geometric space, and whose scores determined the position of the representation in the space. The relative strength of associations between words were measured by calculating the distance between objects in geometric space.

This idea that the strength of associations between words is linked to the distance between

their representations in geometric space, was enunciated well by Schütze and Pedersen [1993] in their early work on word spaces:

> *. . . semantically related words are close, unrelated words are distant.*

In contrast to the affective scales (i.e., dimensions) employed by Osgood et al. [1957], the dimensions in a word space often relate to a pre-defined context, such as the frequency of words within a document. For example, Figure 1.1 illustrates an example of a two-dimensional space where the x-axis represents the number of times each word appears in document 1 (D1) and the y-axis represents the number of times each word appears in document 2 (D2). Based on this word space (Figure 1.1), it can be deduced that both documents contain information about jungle and desert environments, with document 1 having more occurrences of objects found in desert environments and document 2 having more occurrences of objects found in jungle environments.

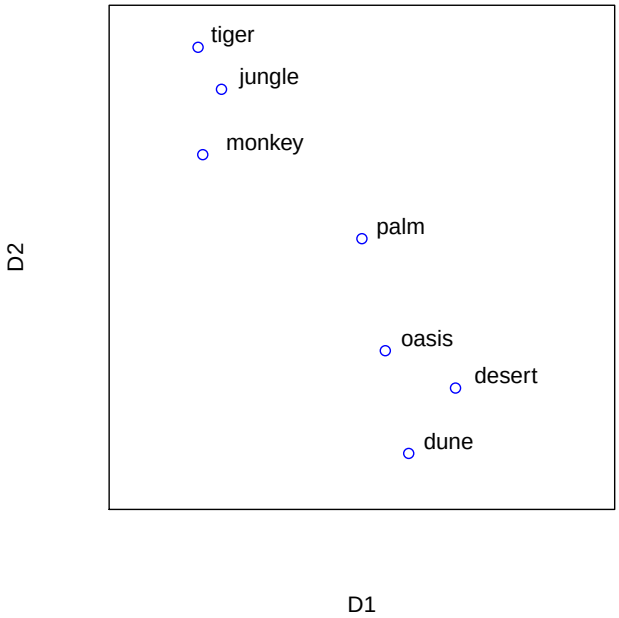

Figure 1.1: Example of a two-dimensional word space.

Models that represent *words* within a multi-dimensional space are often referred to as *semantic space models* (SSMs), due to the distances between words in the space reflecting possible semantic groupings [Lund and Burgess, 1996, Osgood et al., 1957, Padó and Lapata, 2007]. Semantic groups can relate to linguistic classes, such as nouns and verbs, or a more general topic grouping, like natural environments (i.e., desert and jungle) as illustrated in Figure 1.1.

Groupings formed within SSMs have also been shown to support the modelling of syntagmatic and paradigmatic associations found between words in natural language [Rapp, 2002, Schütze and Pedersen, 1993]. Therefore, it is argued that SSMs provide a natural fit for developing computational models of word meaning, grounded on the theories of structural linguistics.

It is acknowledged that co-occurrence patterns of words alone are unlikely to provide a full account of word meaning [French and Labiouse, 2002a, Smolensky and Legendre, 2006]. However, there is a growing body of research demonstrating that psychologically relevant and plausible representations of word meaning can be built from word distributions found in a collection of natural language documents, often referred to as a corpus [Bullinaria and Levy, 2007, Jones and Mewhort, 2007, Landauer and Dumais, 1997, Lund and Burgess, 1996, Schütze, 1993]. These *corpus-based* SSMs are considered to be relatively cheap and flexible as they do not rely on expensive hand-crafted linguistic resources, such as dictionaries or thesauri, commonly used in natural language processing [Dennis and Harrington, 2001, Lin, 1998, Mitchell and Lapata, 2008, Padó and Lapata, 2007].

For these reasons, corpus-based SSMs have seen growing popularity in the field of cognitive science and artificial intelligence [Kanerva et al., 2000, Landauer and Dumais, 1997, Lund and Burgess, 1996, Rohde et al., 2006, Turney and Pantel, 2010] and will be the focus of this research. The following section provides a brief overview of a number of key developments in the evolution of SSMs, and in so doing, highlights a number of gaps in the research that are used to motivate this work.

1.3 The Evolution of Corpus-based Semantic Space Models

Popular corpus-based SSMs, like the *Hyperspace Analogue to Language* (HAL [Lund and Burgess, 1996]) model and *Latent Semantic Analysis* (LSA [Landauer and Dumais, 1997]) represent the occurrence patterns of words found in training documents in high dimensional space. The dimensions within the space often relate to a predefined context within the training corpus. As outlined in the toy example (Figure 1.1), one such context may be the frequency of words within each training document. To illustrate a more realistic scenario, consider a training corpus that contains 200,000 documents. In this case, each word's final representation could be a vector of 200,000 dimensions, whose element values correspond to the frequency count

of that word within each document. This example demonstrates the important point that the dimensions used within an SSM do not not relate to physical dimensions (i.e., 3 dimensional space). In fact, researchers have provided evidence to suggest that the human semantic space has around 300 dimensions [Lowe, 2000].

Working in high dimensional space can have significant computational costs [Beylkin and Mohlenkamp, 2002, 2005]. Therefore, as the size of the training collection increases, so too does the size of the representations, and hence the demand for memory and processing power. This drawback has seen SSM researchers apply mathematical techniques to reduce the dimensionality of the space, as in the well known *latent semantic analysis* (LSA) [Landauer and Dumais, 1997] model, which uses a dimension reduction technique called *singular value decomposition* (SVD) [Golub and Reinsch, 1970]. SVD reduces the semantic space to a small number of latent dimensions (often less than 300), which makes performing similarity judgements much more efficient. However, the SVD process itself is computationally expensive.

The computational complexity associated with working in high-dimensional spaces has led to the development of *fixed dimension* approaches, such as *Random Indexing* (RI) [Kanerva et al., 2000]. Fixed dimension SSMs use approximation techniques to set the dimensionality of the vector representations independent of the number of training documents or words in the vocabulary. These approaches are much more computationally efficient and have demonstrated performance comparable to humans on synonym judgement tasks when vector representations of 1,800 dimensions are used [Karlgren and Sahlgren, 2001].

A common criticism with SSMs that build representations solely from co-occurrence patterns, is that they do not capture structural information present in natural language [French and Labiouse, 2002b, Perfetti, 1998]. This is argued to be responsible for the generation of non-human like errors on some semantic tasks [Perfetti, 1998]. For example, when asked to select the best synonym for *physician*, LSA [Landauer and Dumais, 1997] chose *nurse* over *doctor*. Researchers have since developed SSMs that increase *structural information* within the representations using order-encoding algorithms [Jones and Mewhort, 2007, Sahlgren et al., 2008].

Even with dimension reduction and order-encoding, corpus-based SSMs have failed to demonstrate robust performance across a wide range of semantic tasks. This weakness has seen traditional SSMs labeled as *one model, one task* approaches [Baroni and Lenci, 2010].

Researchers have since argued that capturing co-occurrence information for sequences of words, e.g., phrases, may be required to produce robust performance across a wider range of tasks, including analogy making [Baroni and Lenci, 2010, Turney and Pantel, 2010]. One approach to achieve this involves the use of *higher-order tensor* representations.

A tensor, as used in this work, is a multi-dimensional array [Kolda and Bader, 2009]. More formally, an N-way or Nth-order tensor is an element of the tensor product of N vector spaces, each of which has its own coordinate system. Lower-order tensors relate to vectors (first-order tensors) and matrices (second-order tensors), and within SSMs often hold information about co-occurrence information between individual words. However, third-order tensors can be thought of as an array of matrices, as shown in Figure 1.2 where the matrix formed by the rows of I and columns of J are K deep.

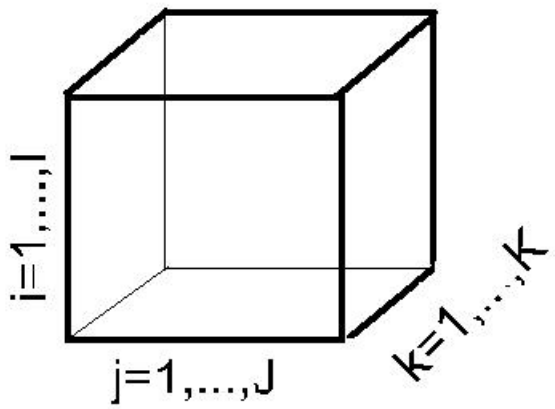

Figure 1.2: A third-order tensor, $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$.

higher-order tensors can be used to capture co-occurrence information of n-tuples [Baroni and Lenci, 2010]. An n-tuple is an ordered list of n elements. This can include n-grams, which are subsequences of n items taken from a given sequence, but extends to include non-adjacent elements from a list. For example, in the sentence, *river to the sea*; the set of 2-tuples that can be formed from this sequence of words is: {*river-to*, *river-the*, *river-sea*, *to-the*, *to-sea*, *the-sea*}.

To illustrate how tensors can be used to store co-occurrence information of 2-tuples and words, consider the co-occurrences of the 2-tuple *river-to* and the word *sea* in our example sentence above. As *river-to* and *sea* co-occur only once, a value of 1 (one) can be stored at element i, j, k in the third-order tensor in Figure 1.2, where i, j, k may correspond to the vocabulary indexes of *river*, *to* and *sea*, respectively.

As SSMs using higher-order tensor representations often incur much higher computational costs [Beylkin and Mohlenkamp, 2002] novel approaches to storage are required to make these computationally tractable. A number of SSMs using higher-order representations have been proposed [Baroni and Lenci, 2010, Humphreys et al., 1989]. Those that have achieved robust performance across a broad range of tasks have often relied on external linguistic resources and still require the entire tensor representations to be stored, which makes them costly and computationally inefficient [Baroni and Lenci, 2010]. Other models using higher-order tensor representations have remained purely theoretical, having never been implemented or evaluated [Humphreys et al., 1989]. In short, higher-order tensor models that have been developed show promising effectiveness across a broader range of tasks, when compared to traditional SSMs, however, they appear to face significant computational challenges.

The evolution of SSMs presented here, has highlighted the benefits of (i) dimension reduction techniques, (ii) order-encoding algorithms and (iii) the use of higher-order tensor representations. It is proposed that the next generation of SSMs will include many of these evolutionary features. However, the evolutionary story would not be complete without providing insights into how information about word associations is accessed within existing SSMs.

Traditionally, corpus-based SSMs have used a single measure of similarity, like a geometric measure (e.g., cosine metric or Euclidean distance) [Bullinaria and Levy, 2007, Rapp, 2002, Sahlgren et al., 2008]. These measures access, what is acknowledged to be, some mix of syntagmatic and paradigmatic information about the representations within the semantic space [Rapp, 2002, Schütze and Pedersen, 1993]. However, no corpus-based SSM appears to formally model and explicitly mix syntagmatic and paradigmatic information in a controlled manner when determining the similarity of words. Given the meaning of words within structural linguistics relies on both types of association, there appears to be a gap in the SSM research, in which a model of word meaning could be developed and evaluated.

Central to this gap is the lack of any formal corpus-based model of word meaning that explicitly models and combines measures of syntagmatic and paradigmatic word associations within a single framework. The idea of explicitly modelling each association for sentence processing has been put forward in the past [Dennis and Harrington, 2001, Jacquemin, 1999], however, these approaches relied on external linguistic resources, i.e., they are not purely corpus-based and hence are outside the scope of this research (Section 2.1). A number of corpus-based models have evaluated methods for measuring syntagmatic and paradigmatic

information [Rapp, 2002, Schütze and Pedersen, 1993], however, it appears that no formal framework to combine them into a model of word meaning has been presented.

Explicit modelling of syntagmatic and paradigmatic associations would allow a controlled mix of information about each type of association to be applied to a given task. This is an attractive feature given some tasks rely more heavily on information about one type of association (e.g., synonym judgements rely more heavily on information about paradigmatic associations). This feature, combined with the the evolutionary changes in SSM technology presented earlier, may provide greater adaptability and robustness within a corpus-based model of word meaning. This discussion leads to the following question: *What improvements in performance and robustness can be achieved by a next generation SSM that explicitly models and combines syntagmatic and paradigmatic information?*

1.4 Applications of SSMs

Semantic technologies, including SSMs, are likely to have transformative powers in addressing the issues caused by the inability of computers to understand human language [Turney and Pantel, 2010]. To illustrate their potential, consider the information seeking process undertaken by a user on the Internet. In most cases, a user has an *information need*, such as: "*What is the best coffee machine to buy for under $500*". Users commonly use short two or three word queries to represent their information need [Carpineto and Romano, 2012], such as "*best coffee machine*", due to difficulties users have in expressing their information need [Cleverdon et al., 1966, Metzler and Croft, 2007]. This query is then used to search the vast collection of web documents on the Internet to determine the most relevant information relating to the query. It is common for a user to refine their query until the search engine retrieves results considered by the user to be more relevant to their information need.

This information seeking process intuitively contains steps that rely heavily on word meanings, and specifically how these meanings are established in the presence of other words, e.g., while viewing a snippet from a query biased summary produced by the search engine. This is illustrated by considering the importance of word meanings when a user transforms their information need into a shortened query for submission to the search engine. In addition, there is also the implicit use of word meanings which may often be overlooked by the user. For example, the user's information need may implicitly involve associations to words

like "*question, lowest, price, taste, reliable, espresso, maker, purchase, local retailer, on-line, cash/eftpos/credit, answer*" to come into mind. Syntagmatic and paradigmatic associations can be argued to exist between these and the original query words (*best, coffee, machine*), including syntagmatic: (*best-price*), (*coffee-taste*), (*machine-espresso*); and paradigmatic: (*best-lowest*), (*coffee-espresso*), (*machine-maker*).

The user's *real information need* may actually be something like: "*I have a question; what is the most affordable/(lowest price), coffee/espresso machine/device/unit that produces great tasting coffee, (hardly breaks down)/(is reliable), costs less than $500, that I can purchase online or from a local retailer? In the answer please specify what types of payment the vendor accepts (i.e., cash/eftpos/credit)*".

This example demonstrates the well known fact, that a query is often a very imprecise representation of a user's *real* information need [Cleverdon et al., 1966, Metzler and Croft, 2007]. Current state-of-the-art techniques for expanding a query representation are based on statistical techniques that primarily model syntagmatic associations between words [Symonds et al., 2012a]. Given the reliance of the user's real information need on word meanings, which in structural linguistics are based on syntagmatic and paradigmatic associations, and the fact that current query expansion techniques rely primarily on syntagmatic associations, then the intuitive question becomes: *Can a corpus-based model of word meaning, using both syntagmatic and paradigmatic associations, be used to underpin a query expansion technique that provides significant improvements in information retrieval effectiveness?*

1.5 Hypotheses and Research Questions

Addressing the language barrier between humans and computers would allow the potential benefits of computers to be more effectively realized. Models of word meaning have the potential to augment information processing technologies with word meanings that align with humans and thus offer the prospect of a more harmonious interaction between human and computer. Corpus-based SSMs provide a low cost semantic technology for building representations of word meaning and have great potential to transform this problem.

1.5.1 Research Question 1

The evolution of SSMs suggests that the next generation of models will likely include the use of: (i) dimensionality reduction techniques, (ii) structural encoding approaches, and (iii) higher-order tensor representations. These will allow SSMs to remain computationally efficient while achieving robust performance on a wide range of tasks. This likelihood, along with the lack of research into explicit modelling of syntagmatic and paradigmatic information leads to the hypothesis that: *more effective performance on a wider range of semantic tasks can be achieved by next generation SSMs that use a formal framework to explicitly model syntagmatic and paradigmatic information*.

The research question to test this hypotheses becomes:

1. Can a corpus-based model of word meaning, formally combining syntagmatic and paradigmatic information, provide superior performance on semantic tasks when compared to existing corpus-based SSMs?

This thesis responds to this question in the affirmative, through (i) a detailed review of existing corpus-based SSMs (Chapter 2), (ii) the development of a new, efficient and effective corpus-based SSM called the *Tensor Encoding* (TE) model (Chapter 3), and (iii) the evaluation of the TE model against a number of benchmark models on a synonym judgement, semantic distance, semantic categorization and medical concept similarity judgement task. The results are published within a number of national and international peer-reviewed proceedings [Symonds et al., 2011a, 2012b,c] (Chapter 4).

This research also outlines the development of a new tensor compression technique, known as *tensor memory compression* (TMC), which enables the TE model to work efficiently with high-order tensor representations.

1.5.2 Research Question 2

Given the dependence on word meanings when a user formulates their query within the information search process, information retrieval provides an intuitive task on which to apply advances in SSM technology. As current techniques for augmenting query representations within the information retrieval process rely primarily on only half the associations underpinning word meanings (i.e., syntagmatic associations), a resulting hypothesis becomes: *using a*

corpus-based model of word meaning, that formally synthesise syntagmatic and paradigmatic information, should allow query representations to be augmented to be more like the user's real information need and lead to significant improvements in retrieval effectiveness. The research question to test this second hypothesis becomes:

2. Can a corpus-based model of word meaning, formally synthesising syntagmatic and paradigmatic information, be used to augment query representations and provide significant improvements in retrieval effectiveness over current information retrieval models?

Again, this thesis responds to this question in the affirmative, through (i) a detailed review of current query expansion techniques (Chapter 5,) the development of a new query expansion technique, known as *tensor query expansion* (TQE) (Chapter 6), and (iii) the evaluation of TQE against a number of benchmark models across a wide variety of industry accepted data sets for the task of ad hoc retrieval (Chapter 7). The results are published within two national peer-reviewed proceedings [Symonds et al., 2011b, 2012a] and an international proceedings [Symonds et al., 2013]. This research discovers that optimal effectiveness is achieved when both syntagmatic and paradigmatic information are used to augment query representations within the information retrieval process.

Concluding remarks relating to this thesis and proposals for future work are outlined in Chapter 8. The areas of potential application of the TE model are extensive, and include applications outside of linguistics.

Part I

Theory

Chapter 2

Semantic Space Models

2.1 Scope and Motivation

This chapter provides a review of SSM literature, specifically relating to corpus-based SSMs that do not rely on external linguistic resources, such as hand-crafted knowledge bases, parsers or *part-of-speech* (POS) taggers. Computational models that rely on linguistic resources often incur a cost relating to the creation and maintenance of these resources, and are language specific [Dennis and Harrington, 2001, Jacquemin, 1999, Lin, 1998, Mitchell and Lapata, 2008, Padó and Lapata, 2007]. By not relying on external linguistic resources, this cost is avoided and a less cumbersome model is more likely achieved.

In addition, modelling linguistic relationships based on word occurrence patterns in natural language (i.e., corpus-based) allows more *contextual* representations of words to be formed. Context-sensitive representations will more naturally support the aspects of meaning that vary with time and individual interpretation (Section 2.2). This is in contrast to the often rigid relationships defined within external linguistic resources, like dictionaries and thesauri.

On the other hand, restricting the scope of this work to corpus-based SSMs that are independent of linguistic resources, means a full account of human lexical capabilities is unlikely to be achieved [French and Labiouse, 2002a, Smolensky and Legendre, 2006], as these approaches would have difficulty modelling homographs [Schütze, 1998]; which are words with the same form but different meanings, as illustrated by the word *bank* in the following sentences: (i) *The* ***bank*** *was robbed*; (ii) *The water lapped against the* ***bank*** *of the river*.

However, there is a growing body of research demonstrating that psychologically relevant and plausible representations of word meaning can be created solely from the occurrence patterns of words in natural language [Bullinaria and Levy, 2007, Jones and Mewhort, 2007, Landauer and Dumais, 1997, Lund and Burgess, 1996, Schütze, 1993]. The central assumption has been that the meaning of a word can be deduced from the distribution of the surrounding words [Harris, 1968]. This assumption is supported by research showing language development in young children relies heavily on these occurrence patterns when building word meanings [Jones and Mewhort, 2007].

For these reasons, the focus of this work is on corpus-based SSMs that are independent of external linguistic resources, and the experiments carried out will compare models within this class. However, SSMs outside this group may be discussed if they display relevant features, not specific to their use of linguistic resources, that provide insight into research decisions made in this work.

To carry out a rigorous review of corpus-based SSMs and how they may be used to model word meaning, as defined by the differential view of meaning, a sound understanding of structural linguistic theories is required.

2.2 Structural Linguistic Theory

Early linguists identified meaning as one of the slipperiest tools in their methodological kit [Newman, 1952]. However, structural linguistics has been argued to provide a relatively clean theory of word meaning, free of psychology, sociology and anthropology [Holland, 1992].

The key ideas within structural linguistics are credited to the Swiss linguistic Ferdinand de Saussure (1916). Saussure's posthumously published work, *Course in General Linguistics* based on lecture notes, is widely regarded as the most influential work in linguistics, and he as the founder of modern linguistics [Lyons, 1968]. Even though these lecture notes were brought together and edited by others, and Saussure did not explicitly use the term *structuralism* in his notes, the "differential view of meaning" [Elffers, 2008, Pavel, 2001, Sahlgren, 2006], referred to in this work, relates to the structural study of language proposed by Saussure.

Structural linguistics considers language as an ordered sequence of *signs*. Within semiotics, which is the study of signs, a sign is defined as:

> any physical *form* that can be imagined or made externally (through some physical medium) to stand for an object, event, feeling, etc., known as a *referent*, or for a class of similar (or related) objects, events, feelings, etc., known as a *referential domain* [Sebeok, 2001].

For Saussure, a sign exhibits a dyadic property, in that it is made up of two components, the *signifier* (something physical, i.e., sound, letters, gestures, etc.) and its *signified* (an image or concept to which the signifier refers), as depicted by the dyadic sign in Figure 2.1.

The relationship between signifier and signified is argued by Saussure to be arbitrarily assigned, as there is no evident reason why a signifier, like the letters that make up the word *tree*, would be associated with the signified, i.e., the concept of "an arboreal plant". Any well-formed signifier could have been used. A well-formed signifier is one that meets the structural constraints of a language, e.g., *tree* is well-formed in English, but *tkdf* is not.

Saussure noted that within a society the relationship for a given signifier may vary, due to personal interpretation. Even within an individual, the relationship between the signifier and signified can change over time, because of a change in understanding. This temporal property of signs led Saussure to divide the study of signs into two branches, *synchronic* and *diachronic*. The former studying signs at a given point in time, and the latter focusing on how signs change over time.

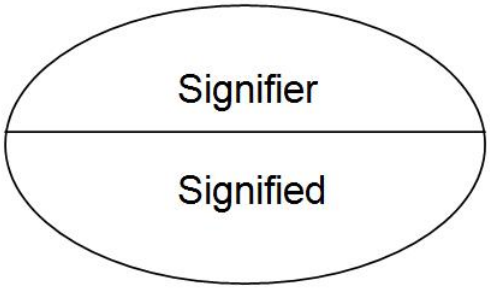

Figure 2.1: A dyadic sign.

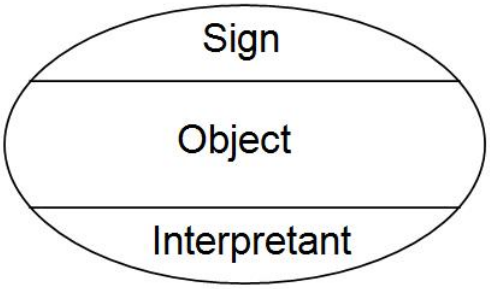

Figure 2.2: A triadic sign.

A different approach to representing this temporal property of signs was proposed by American philosopher Charles S. Peirce (1839-1914), who proposed that a *sign* depends on an *object* in a way that enables an *interpretation*, as depicted by the triadic sign in Figure 2.2 [Sebeok, 2001]. In this representation, the sign is equivalent to Saussure's signifier, an object can be anything thinkable that the sign represents or *encodes*, and the interpretation is the meaning (or *decoding*) of the sign. This representation more directly encapsulates the temporal nature of

signs.

Within structural linguistics, the meaning of a sign is determined from the differences it has with other signs, and hence cannot be considered in isolation to the language. This differential view of meaning was a significant point of departure from previous theories of linguistics at the time. The two fundamental relationships formed between signs in structural linguistics are known as: (i) syntagmatic and (ii) paradigmatic associations.

As defined earlier (Section 1.1), a syntagmatic association exists between two words if they co-occur more frequently than expected from chance within natural language [Lyons, 1968]. Some typical examples may include "coffee-drink" and "sun-hot". These associations can also have varying strengths. Consider the example sentence "A dog bit the mailman". The term "dog" would likely have a stronger syntagmatic association with "bit" than "mailman", based on the fact that the word "bit" would likely co-occur with "dog" more often.

A paradigmatic association exists between two words if they can substitute for one another in a sentence without affecting the acceptability of the sentence [Lyons, 1968]. The "*acceptability* of a sentence" is acknowledged to be ambiguous, and hints at the probabilistic nature of the relationship [Haas, 1973]. Typical examples include the word pairs "drink-eat" and "quick-fast" [Rapp, 2002]. In the example sentence, "the dog bit the mailman", the word "bit" could be replaced with "chased", hence "bit and "chased" could be said to have a paradigmatic association. Words sharing a paradigmatic association are often the same part of speech, i.e., related nouns or related verbs.

Syntagmatic and paradigmatic associations exist between linguistic units in language, including phonemes (similar sounds, *cat, hat, bat*), morphemes (morphological variations of words, *run, runner, running*), lexical categories (*noun, verb, adjective*), noun phrases and verb phrases, which allow models grounded in these theories to be applied at different levels of language.

The functional role of these associations have underpinned successful linguistic theories, including those on *sense relations* by Lyons [1968]. However, within this dissertation, a much more computational, rather than functional, approach to leveraging the existence of syntagmatic and paradigmatic associations within language is taken. Therefore, discussion of various linguistic theories derived from Saussure's ideas are restricted, with much more focus being given to the effective modelling of syntagmatic and paradigmatic associations, based on their formal

definitions (Section 1.1), and the observed distributional information found in natural language.

This restricted focus is consistent with past research on corpus-based SSMs that operationalise the distributional hypothesis (Section 1.1) by modelling syntagmatic and paradigmatic associations [Rapp, 2002, Schütze and Pedersen, 1993]. Even though SSM research has highlighted the existence of syntagmatic and paradigmatic associations it often omits any explicit reference to *structural linguistics*.

This omission may be due to the stigma associated with structural linguistics, created by strong, critical views aired by linguistic luminaries, like Noam Chomsky and John Lyons, relating to its inability to provide a complete model of language [Haas, 1973]. However, given theories of generative grammar have also been unable to achieve such lofty heights, there appears to be no compelling reason to omit ideas from structural linguistics.

The fact that structural linguistics is rarely highlighted in SSM literature, may explain why no corpus-based model of word meaning has explicitly modelled and combined measures of syntagmatic and paradigmatic associations. Several approaches have explicitly modelled these associations [Rapp, 2002, Schütze and Pedersen, 1993, Sitbon et al., 2008], however, there does not appear to be any corpus-based examples where these associations are explicitly combined to create a formal model of word meaning. This highlights a gap in the research that motivates a computational approach to modelling and combining measures of syntagmatic and paradigmatic associations that could underpin a formal model of word meaning.

2.3 Mathematical Framework

Modelling semantic relationships using word occurrence patterns has been achieved using both probabilistic [Lee, 1999, Lin, 1998, Yuret, 1998] and geometric approaches [Landauer and Dumais, 1997, Lund and Burgess, 1996, Turney and Pantel, 2010]. Using geometric representations does not exclude the use of probabilistic methods, and recent work with SSMs has demonstrated that improved task effectiveness can be achieved when both frameworks are combined [Bullinaria and Levy, 2007].

An intuitive motivation for using geometric representations comes from the finding that relationships between words are stored spatially within the brain. This was observed during brain activation research performed using *functional magnetic resonance imaging* (fMRI) technology

on human subjects undergoing word priming experiments [Mitchell et al., 2008].

Perhaps a more practical advantage of using geometric representations comes from the ability to use the powerful tools of linear algebra to confer non-commutative properties observed in natural language [Meyer, 2000]. For example, the Kronecker product ($\otimes$) of two vectors a_m and b_n, also known as the outer product, is defined as:

$$a \otimes b^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} (b_1 \quad b_2 \quad \dots b_n) = \begin{pmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & a_2b_2 & \dots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_mb_1 & a_mb_2 & \dots & a_mb_n \end{pmatrix},$$

where b^T is the transpose of vector b.

The Kronecker product is a generalisation of the outer product of vectors, and can be applied to higher-order tensors [Meyer, 2000]. For example the Kronecker product of matrix $A_{m\times n}$ and $B_{p\times q}$ is defined as:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}.$$

The Kronecker operator exhibits non-commutative properties as $A \otimes B \neq B \otimes A$. If A and B are the tensor representations of two unique words in a vocabulary, then the two different results produced by $A \otimes B$ and $B \otimes A$ can be argued to be analogous to the various meanings produced by combining words in a different order. To illustrate, consider how the meaning changes when the order of the words *space* and *program* are swapped. *Space program* often relates to a country's plans for space exploration, whereas, *program space* refers to the area in computing memory where the instructions for a computer program are normally stored.

The above points highlight some of the benefits gained from using geometric settings to model relationships between words. However, many corpus-based SSMs use geometric representations that contain explicit frequency counts of words, or word co-occurrence frequencies.

Therefore, opportunities for incorporating information theoretic measures, which are set within a probabilistic frameworks, also exist. This is important, as recent SSM research has demonstrated that superior task effectiveness can be achieved when both mathematical frameworks are combined [Bullinaria and Levy, 2007].

2.4 Geometric Representations

Representing occurrence patterns of words in geometric objects is often achieved by collecting word frequencies and placing them within high-dimensional *context* vectors [Turney and Pantel, 2010]. Within this research, the *context vector* will refer to the final vector representation of a word that is used in similarity calculations within vector based SSMs.

The method of creating the context vectors determines the associated computational cost of the model, in terms of memory requirements, time to build the space and time to perform similarity calculations. This process also determines the type of information that can be extracted, which influences the type of tasks the model can be applied to. For instance, storing the frequency of words in a document within the context vectors, as was done in the example word-space in Chapter 1 (Figure 1.1), may be well suited to document retrieval tasks where documents relating to a specific topic are often sought. However, this document-based context may not be the most effective approach for calculating the semantic relatedness of words.

Within an SSM, *semantically related* words are considered close in geometric distance and unrelated words further apart [Schütze, 1993]. However, across SSM literature more specific types of similarity have been studied, including: *relational similarity*, *attributional similarity* and *semantic similarity*, and an informative discussion of these is provided in Turney and Pantel [2010]. Within this dissertation, the use of the term *similarity* refers to the comparison of the meaning of words, as defined by the differential view of meaning and represented by their mix of syntagmatic and paradigmatic information within the TE model.

The measures used to determine the similarity of representations are often referred to as similarity measures. Often the most effective measure of similarity will differ depending on the task being undertaken, or the type of similarity being evaluated. Example geometric-based similarity measures include the popular cosine metric and Minkowski norms (i.e., city block and Euclidean distance, etc.) [Meyer, 2000]. More detailed discussion regarding the role of similarity measures is provided in Section 3.2.

To better understand the many approaches to storing occurrence patterns of words within context vectors, a review of a number of models that have played an important role in the evolution of SSMs is presented. The models are presented in relation to the evolutionary step which they incorporated to overcome weaknesses in previous SSM design.

2.5 Dimension Reduction and Fixed Dimension SSMs

Within SSM research, the elements making up the vocabulary created from a set of training documents are often referred to as vocabulary *terms*, as opposed to vocabulary *words* as they often include abbreviations, numbers and other compound alphanumerics. The following review of SSM literature will use *word* and *term* interchangeably to refer to the elements that make up the vocabulary. This is done to help improve the readability of the review.

2.5.1 Very High Dimensional Spaces

The process of collecting occurrence patterns of words from large collections of training documents results in very sparse and high dimensional context vectors. These are created because the context information for each word needs to be stored. Two popular contexts used by corpus-based SSMs [1]:

1. **Document context**: stores the distributional information of words found in each document in the training corpus. If there are 200,000 documents in the training corpus and the corpus contains 300,000 vocabulary terms, then to determine the similarity between terms, the model would require 300,000 context vectors, each of 200,000 dimensions.

2. **Word co-occurrence context**: stores the co-occurrence frequencies between each vocabulary term. Therefore, using the first example in which the vocabulary size was 300,000, the model would require 300,000 context vectors each of 300,000 dimensions.

In both cases, the issues associated with analyzing and organising data in these high dimensional spaces, known as *the curse of dimensionality* [Bellman, 1961], results in a very large memory footprint that demands increasingly large amounts of processing time to calculate the similarity of the terms (i.e., as represented by their context vectors).

[1] The Pair-Pattern context outlined in Turney and Pantel [2010] is often found in SSMs that rely on linguistic resources (parsers, POS taggers) and hence is left out of this discussion.

2.5.2 Data Sparseness

In both the document and word co-occurrence contexts above, the context vectors contain many elements with a zero value. This sparseness is always present, as each document contains only a small subset of the vocabulary words, and words within natural language are normally only seen in the presence of a relatively small group of words.

Two popular SSMs that use different approaches to address the curse of dimensionality and the data sparseness problem are the *Hyperspace Analogue to Language* (HAL) and *Latent Semantic Analysis* (LSA) models.

2.5.3 Hyperspace Analogue to Language (HAL)

HAL [Burgess et al., 1998] uses a word co-occurrence context (Section 2.5.1) and builds its context vectors through a two step process. Firstly, a word co-occurrence matrix is created by storing the proximity scaled co-occurrence frequencies of words seen within a sliding context window that is moved across the text of a set of training documents. The resulting HAL matrix is n-by-n, where n is the number of words in the vocabulary, and is made up of rows containing succeeding word co-occurrence frequencies and columns with preceding word co-occurrence frequencies. The co-occurrence frequencies are weighted inversely proportional to the distance between the words found in the training corpus.

To illustrate, consider the HAL matrix shown in Table 2.1, which is created for the toy sentence: *A dog bit the mailman*, using a sliding context window of length 5 (i.e., 2 words either side of the focus word). The co-occurrence information preceding and following each word are recorded separately by the row and column vectors. The values assigned to each co-occurrence are scaled by their distance from the focus word, with words next to the focus word given a value of 2 (when a context window length of 5 is used), and those at the edge of the window scaled by 1.

The final context vectors for each word are formed by concatenating the row-column pairs (i.e., preceding and following co-occurrences). These representations produce a very high dimensional space, twice that of the vocabulary size, within which words that co-occur with similar words will be closely located. These very large representations incur significant computational costs in terms of memory and processing time due to the curse of dimensionality (Section 2.5.1) and data sparseness (Section 2.5.2) issues.

	a	dog	bit	the
dog	2	0	0	0
bit	1	2	0	0
the	0	1	2	0
mailman	0	0	1	2

Table 2.1: Example HAL Space.

To overcome these issues, developers of HAL introduced a compression technique based on discarding elements within the context vectors that had low variance across the row and column vectors. An alternate approach, with a similar effect, is to keep only elements from the top most frequent terms in the vocabulary. This often causes a small drop in task performance, but the computational savings within the model are argued to outweigh the loss in effectiveness [Bullinaria and Levy, 2007].

These dimension compression techniques impact on the model's ability to make similarity judgements between low frequency words, as no co-occurrence information between the low frequency test words would exist. This may require the compression technique to be modified to keep elements containing co-occurrence frequency between any test words. This can be an issue when working with unseen test data.

Another drawback of HAL based approaches stems from the relatively small context windows (< 10) often used to build the context vectors [Burgess et al., 1998]. Capturing only near neighbour co-occurrences leads to the criticism that HAL based approaches do not include strong associations found between words further apart in language, which is argued to be contrary to the way human memory has been shown to function [Perfetti, 1998].

Inspired by HAL, Rohde et al. [2006] constructed the *Correlated Occurrence Analogue to Lexical Semantic* (COALS) model that demonstrated robust performance across a wide variety of natural language based cognitive tasks. The performance of the COALS model was compared to a number of high performing benchmark systems, including WordNet. WordNet is a well-known ontological based approach to modeling human lexical semantic knowledge that uses a large network of word forms, their associated senses, and various links between those senses. The COALS study demonstrated that robust cognitive performance can be achieved by HAL-based SSMs.

A weakness of the HAL model, pointed out by the COALS researchers [Rohde et al., 2006] was the effect high frequency columns (high variance columns) had on the distance measures used by geometric models. They employed a normalisation factor to reduce the impact of these high variance columns.

So in summary, HAL captures occurrence patterns of words in the form of scaled co-occurrence frequencies along with some directional information. For large vocabularies, HAL-based approaches often reduce the dimensionality of the space by removing vector elements of low frequency terms, and can experience noise issues from high frequency terms. Unlike HAL's reduction of dimensions through compression, other corpus-based SSMs, such as *latent semantic analysis* (LSA) [Landauer and Dumais, 1997], have applied more formal mathematical techniques to achieve reductions in dimensionality.

2.5.4 Latent Semantic Analysis (LSA)

Mathematicians modelling real world problems using linear algebra have developed a number of techniques for decomposing large matrices into a number of smaller matrices, that can be worked on much more efficiently [Berry et al., 1999]. Researchers working on corpus-based SSMs using document-based context vectors, developed a model known as *latent semantic analysis* (LSA) [Landauer and Dumais, 1997] that used a matrix reduction technique called *singular value decomposition* (SVD). This approach was originally evaluated on information retrieval applications, and was called *latent semantic indexing* (LSI) [Dumais et al., 1988]. However, Within the literature LSA and LSI have become somewhat synonymous. Therefore, within this thesis a corpus-based SSM using SVD will be referred to as LSA.

When dealing with a term-document context, the SVD process that underpins LSA decomposes the term-document matrix into three separate matrices, one of which represented the k most significant latent concepts within the term-document matrix. SVD takes the normalised $m \times n$ matrix X, where m signifies the number of terms in the vocabulary and n the number of documents in the collection, and decomposes it in the following way [Meyer, 2000]:

$$X = USV^T$$

where U is a $m \times m$ orthogonal matrix whose columns define the left singular vectors of X; V is an $n \times n$ orthogonal matrix whose columns define the right singular vectors of X (note V is transposed, so they appear as rows in Figure 2.3); and S is the $m \times n$ diagonal

matrix containing the non-negative singular values of X. If the matrices (U, V and S) are setup so that the singular values are in decreasing order, then by taking the k largest singular values in S, the matrices can be modified to only use the first k columns of U, k rows of V to produce a least squares approximation to X, as shown in Figure 2.3. Working with the reduced dimension approximation of X incurs an error related to value of k but reduces the computational complexity of making similarity measures within this much lower dimensional space [Meyer, 2000].

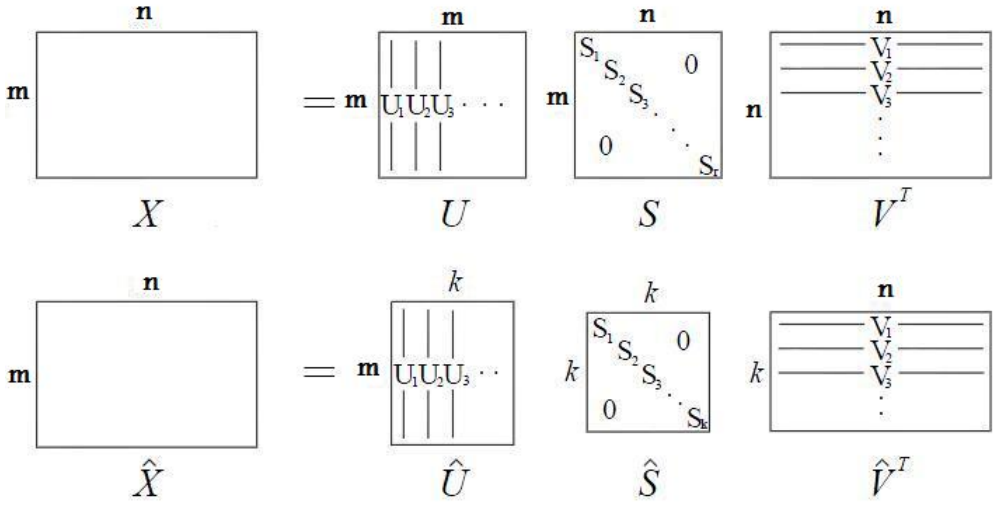

Figure 2.3: The singular value decomposition of matrix X. $\hat{X}$ is the best rank k approximation to X, in terms of least squares [Rohde et al., 2006].

Dumais et al. [1988] outlined how the decomposition process modifies the term-document matrix to reposition the term vectors so that terms used in similar contexts will be closer in vector space. Other researchers have suggested SVD is another method for smoothing the data, that has greatest benefits for sparse data sets [Rapp, 2002].

LSA has been successful at emulating psychological data and scored 64.5% on the *Test Of English as a Foreign Language* (TOEFL) [Landauer and Dumais, 1997]. This score was the same as that achieved by a large sample of students, from non-English speaking countries, that applied for college entrance in the United States. The TOEFL is a standardized test employed by many universities worldwide to evaluate a foreign applicants' knowledge of the English language. In the synonym-finding part of the test, the participant is asked to find the synonyms to certain words. Each question involves the participant choosing one of the four provided words as the synonym for the question word.

Perfetti [1998] argues that even though this performance by LSA on the TOEFL can be

shown statistically, when the errors are inspected more closely, the error patterns between LSA and the students are not closely correlated ($r = 0.44$). For example, for a stem of *physician*, LSA chose *nurse* over *doctor*. Such an error is not expected from a student. This would indicate an in-kind error, in Perfetti's words, and suggests LSA does not truly operate like a human.

Another drawback of LSA is often attributed to the computational costs associated with the SVD process, as well as the need to compose the initial co-occurrence matrix. These computational overheads make the application of LSA on large corpora of electronic documents less attractive.

Another approach that addresses the need to reduce the dimensions and sparseness of SSMs involves the use of approximation techniques in building spaces that hold information about word associations. Within this dissertation, these approaches are referred to as *fixed dimension* SSMs, as they fix the length of the final context vectors independent of the number of documents in the training collection, or in the case of the co-occurrence context, the number of words in the vocabulary.

2.5.5 Random Indexing (RI)

Random indexing (RI) [Kanerva et al., 2000] was motivated by observations made by researchers who demonstrated that there are many more nearly orthogonal directions in a high-dimensional space [Hecht-Nielsen, 1994]. This means that orthogonality within a space can be approximated by selecting random directions in a high-dimensional space.

By assigning each word in the vocabulary a random vector in a high dimensional space (known as an environment vector), the context vectors for each word can be created by summing the environment vectors of words that appear within the sliding context window.

A more detailed explanation of the two step process used by RI is provided by Sahlgren [2005]:

- "Firstly, each context (e.g. each document or each word) in the data is assigned a unique and randomly generated representation, call this an environment vector. These environment vectors are sparse, high-dimensional, and ternary, which means that their dimensionality (d) is on the order of thousands, and that they consist of a small number (often less than 10) of randomly distributed $+1$s and -1s, with the rest of the elements of the vectors set to

0.

- Secondly, context vectors are produced by scanning through the text, and each time a word occurs in a context (e.g. in a document, or within a sliding context window), that context's d-dimensional environment vector is added to the context vector for the word in question. Words are thus represented by d-dimensional context vectors that are effectively the sum of the words' contexts."

When computed across the entire training corpus using a context window this has the effect of reducing the geometric distance between term vectors of words with similar meanings, as synonyms or words used in the same context will keep similar word company. Due to the random nature of the environment vectors the evaluation of an RI model is done by averaging the results over a number of trials.

A brief evaluation showed that RI was able to produce similar results to LSA on TOEFL [Kanerva et al., 2000]. It was noted by Gorman and Curran [2006] that frequency weights needed to be added to the model to improve performance on large data sets. RI has also been evaluated within cross language information retrieval [Sahlgren and Karlgren, 2002]. No significant improvements in retrieval effectiveness were reported.

2.5.6 Summary

Critics of semantic space models argue that semantic models based on co-occurrence statistics alone will not be sufficient to build reliable lexical representations [French and Labiouse, 2002b]. They believe further computational apparatus will be required. Glenberg and Robertson [2000] highlight the fact that semantic space models using abstract symbols such as co-occurrence statistics are unable to truly know the meanings of words and hence cannot form an adequate basis for human meaning. This may indicate that stronger forms of syntactic information need to be bound into semantic models.

Even though models based on HAL, LSA and RI have been shown to simulate human performance on a number of cognitive tasks, it has been argued by Perfetti [1998] that these models do not capture concepts such as syntax or achieve other basic cognitive language abilities. A relevant example, includes the fact that LSA chose *nurse* over *doctor* when asked to determine the closest match to *physician* in a synonym judgement test. The lack of word order

information in LSA is a result of the way in which it builds its context vectors, however, even though HAL would appear to hold word order information it does not formally encode order information [Jones and Mewhort, 2007]. Recent research suggests that encoding word order information can assist with addressing these limitations and is more likely to match the trends found in human data [Jones and Mewhort, 2007].

2.6 Encoding Structure

Two recent approaches that have encoded word order within the representations include the *Bound Encoding of the Aggregate Language Environment* (BEAGLE) model [Jones and Mewhort, 2007] and the *permuted random indexing* (PRI) model [Sahlgren et al., 2008]. Both the BEAGLE and PRI models are fixed dimension approaches that use approximation techniques, like RI.

2.6.1 Bound Encoding of the Aggregate Language Environment (BEAGLE)

Like RI, the BEAGLE model [Jones and Mewhort, 2007] is built from fixed dimension, randomly assigned, environment vectors. However, these environment vectors ($\vec{e}_i$) are formed by sampling from a Gaussian distribution with $\mu = 0$ and $\sigma = 1/\sqrt{d}$, where d is the dimension of the vectors. Resulting in dense, rather than sparse, environment vectors.

A second point of departure from the RI approach comes from the fact that BEAGLE builds up two separate representations for each word, a *context* and an *order* vector, which are summed at the end of the vocabulary building process to form what is referred to as a *memory vector*. The context vector stores associational information (like RI) and the order vector is formed using a unique binding operation based on a compressed outer product operation that encodes word order. As outlined in Section 2.3, the outer product is the operator name given to the Kronecker product operator when working with vectors.

As BEAGLE was the inspiration for a number of design choices made in developing the TE model (Chapter 3), a more detailed description of the workings behind BEAGLE will be provided. The process for creating the context and order vectors includes:

1. The context vector is built by summing the environment vectors of all other terms in the set C, formed by the context window (e.g., the context of a sentence was used in the

original BEAGLE model) with the existing context vector:

$$\vec{c}_w = \vec{c}_w + \sum_{k \in \{C | k \neq w\}} \vec{e}_k, \tag{2.1}$$

where $\vec{e}_k$ is the environment vector of term k in the sentence. For the set of ordered term, C that make up the sentence, "*a dog bit the mailman*", the context vector for dog ($c_w = c_{\text{dog}}$ in Equation (2.1)), would be:

$$\vec{c}_{dog} = \vec{c}_{dog} + \vec{e}_{bit} + \vec{e}_{mailman}.$$

Note, stop words (i.e., "a", "the") are ignored in creating the context vector.

2. The order vector is built by performing a binding process with words found within a λ-length window moved across each sentence, such that:

$$\vec{o}_w = \vec{o}_w + \sum_{l=1}^{p\lambda-(p^2-p)-1} bind_{wl},$$

where p is the position of the word in the sentence, and $bind_{wl}$ is the lth convolution binding for the word being coded. To demonstrate what the lth convolution binding process is, consider the binding operations for the word *dog* in the sentence: "*a dog bit the mailman*":

$$\left.\begin{aligned} bind_{dog,1} &= e_a \circledast \Phi \\ bind_{dog,2} &= \Phi \circledast e_{bit} \end{aligned}\right\} Bigrams$$

$$\left.\begin{aligned} bind_{dog,3} &= e_a \circledast \Phi \circledast e_{bit} \\ bind_{dog,4} &= \Phi \circledast e_{bit} \circledast e_{the} \end{aligned}\right\} Trigrams$$

$$\left.\begin{aligned} bind_{dog,5} &= e_a \circledast \Phi \circledast e_{bit} \circledast e_{the} \\ bind_{dog,6} &= \Phi \circledast e_{bit} \circledast e_{the} \circledast e_{mailman} \end{aligned}\right\} Quadgrams$$

$$bind_{dog,7} = e_a \circledast \Phi \circledast e_{bit} \circledast e_{the} \circledast e_{mailman} \} Tetragram$$

where $\vec{\Phi}$ is a constant random placeholder vector (constant across all terms) that is sampled from the same element distribution as the environment vectors, and $\circledast$ is a modified *circular convolution* that produces Holographic Reduced Representations (HRRs) [Plate,

2001]. HRRs have been used recently to underpin a large scale model of the functioning brain [Eliasmith et al., 2012], as this type of compression is argued to be analogous to the way in which the human brain stores information.

Circular convolution ($\odot$), depicted in Figure 2.4, uses an algorithm of multiplication that allows the outer product of environment vectors to collapse to a vector with the same dimensions. To ensure the binding process is non-commutative, such that $a \circledast b$ and $b \circledast a$ produce unique associations, the environment vectors are permuted before circular convolution is performed. Permuting involves swapping the position of element values within environment vectors. For BEAGLE, the permutation function used was dependent on the position of the word in the binding process.

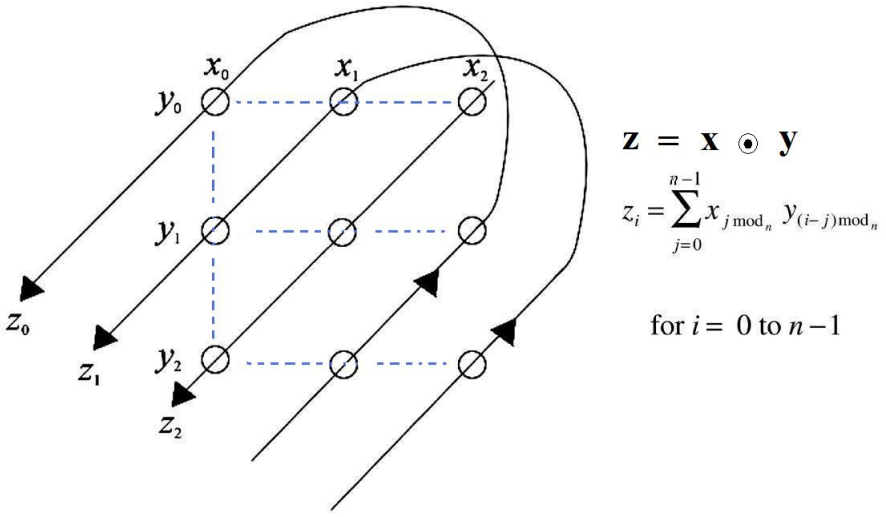

Figure 2.4: Circular convolution represented as a compressed outer product of 3-dimensional vectors x and y. Vector z is the resulting compressed 3-dimensional vector. The values i and j represent the row and column indices, respectively, for an element in the outer-product n-n matrix, where $n = 3$.

The resulting representation for a word is built by summing the context and order vectors into the memory vector, namely:

$$\vec{m}_w = \vec{c}_w + \vec{o}_w. \tag{2.2}$$

Measuring Similarity

To evaluate the effectiveness of BEAGLE in accessing the context and order information stored within the memory vectors, a method of decoding and resonance were tested on a word priming

task.

1. Decoding: The decoding process, uses inverse circular convolution ($\odot$), also known as circular correlation [2]. The process results in the creation of a noisy version of the environment vector that best fits the sequence of words used in the word priming task. By using a cosine metric on all environment vectors the most activated word can be found. Researchers propose that due to circular correlation only being an approximation for inverse circular convolution this may have resulted in low performance of this approach.

2. Resonance: *Resonance* is the tendency of an object to absorb more energy when the probed frequency matches the object's natural frequency. To determine which lexical entries are highly activated by a position in a sentence, the λ-gram convolutions around the position are summed into a probe vector, using Φ in the blank spot. The resulting probe vector is then compared to the memory vectors of terms looking for resonance (words that have this pattern the most). This demonstrated some effectiveness for sequences where the missing word (in *[italics]*) was strongly resonant, for example: "Martin Luther *[King]*" or "the *[emperor]* penguins have come to their breeding grounds". The results of using BEAGLE in word priming experiments demonstrate its value as a memory model.

Synonym Judgement

For other semantic tasks that require more paradigmatic associations (like synonym tests), a geometric distance based on the inner product (cosine metric) of the memory vectors is used. For example, the researchers used the TOEFL synonym matching task to compare BEAGLE's performance to the other semantic space models.

When BEAGLE was trained on the TASA corpus using only context information and vectors of 2048 dimensions, it answered 55.60% of questions on TOEFL correctly. When using both context and order vectors BEAGLE correctly answered 57.81% of the items. Both of these results reported are lower than that achieved by students applying for college entrance in the United States (64.5%) and LSA when trained on the same corpus (65%). However, they do appear to indicate that the inclusion of word order information can improve performance on this task.

[2] Refer to Jones and Mewhort [2007] for implementation details

Another variable, noted by Sahlgren et al. [2008], that may have influenced the effectiveness of the BEAGLE model on the TOEFL synonym judgement task relates to the wide context window used in their experiment (i.e., the length of the sentence). Other research has previously shown that for synonym judgements, which make use of information about paradigmatic associations between words, a narrower context window (i.e., less than 3) improves task effectiveness [Bullinaria and Levy, 2007, Landauer and Dumais, 1997, Sahlgren et al., 2008]. This is due to paradigmatic associations being formed between words that share the same close neighbours within natural language.

Summary

Two significant drawbacks of the BEAGLE model appear to be the computational complexity and noise associated with the binding and decoding processes, respectively. This is likely due to the compression associated with circular convolution, leading to noise existing in the holographic reduced representations.

Another experimental drawback of models using random assignment of environment vectors relates to the instability of the results. This means that evaluating these models robustly requires a number of test runs to be performed and averaged to produce a final effectiveness score.

The second example of a corpus-based SSM that attempts to encode more structural information within representations involves an extension of the RI model.

2.6.2 Permuted Random Indexing (PRI)

Following the work on BEAGLE, Sahlgren et al. [2008] extended the RI model to encode word order. The final representations within this *permuted random indexing* (PRI) model were formed by summing the context and order vectors after the vocabulary building process was complete. The context vectors were created as in RI, however, order vectors were created by summing *permuted* environment vectors. A different permutation operation was performed depending on the position of the word with respect to the target word in the context window. For example, the direction of permutation was based on the position of the word from the target word (say negative for preceding words and positive for following words), and the number of times an environment vector was permuted based on its distance from the target word.

This method of encoding word order is more computationally efficient than BEAGLE's

binding process and when evaluated on the synonym detection part of TOEFL using vectors of 30,000 dimensions was able to score 80%. This research showed that permuting based on the direction alone (i.e., before or after, no consideration of n) was more effective than permuting using proximity and directional information. The performance of the PRI model appeared to display substantial sensitivity to the frequency cut-off used in building the vocabulary, with the score dropping to 45% when no frequency cut-offs were used. This will be an important consideration when evaluating the PRI model on experiments performed later in this work.

2.6.3 Summary

These order-encoding models attempt to address the issue raised by researchers like Perfetti [1998], relating to the lack of structural information used by early SSMs, like LSA. Other approaches to capturing word order information within SSMs have been presented in the past, however, these approaches rely on linguistic resources to achieve this and hence are not considered in the scope of this research [Mitchell and Lapata, 2008, Padó and Lapata, 2007].

It is worth noting that within most corpus-based SSMs all co-occurrences (i.e., even those with low information terms) are considered to have equal importance, i.e., the operation used to add or bind environment vectors treat all co-occurrences in the same way. This insensitivity to the information content of a co-occurrence may also add to the noise in the model and hence explain the importance of applying frequency cut-offs to achieve reasonable effectiveness [Sahlgren et al., 2008].

This observation may be confirmed by the need to use relatively high frequency cut-offs within existing corpus-based SSMs to achieve improved task effectiveness [Rohde et al., 2006, Sahlgren et al., 2008]. Frequency cut-offs determine what terms will exist in the vocabulary, effectively reducing the size of the vocabulary and the existence low information terms. However, for tasks using unseen data, this process may remove test terms, and hence reduce the flexibility of a model.

Researchers have argued that the flexibility of existing SSMs has been limited, with a *one task, one model* approach often being adopted [Baroni and Lenci, 2010, Turney, 2008]. The ability for SSMs to be more flexible is argued to come from new contexts, or the creation of higher-order models that can capture co-occurrence information between multi-word concepts, also known as n-tuples ($n > 1$) [Baroni and Lenci, 2010, Turney and Pantel, 2010]. n-tuples are

formed from considering a number of words as a single concept, as found in phrases like *space program*. The ability to store co-occurrence information about phrases often involves the use of higher-order tensor representations, which have traditionally been thought to involve significant increases in computational costs for a model.

2.7 High-order Tensor Representations

Tensors include the set of scalar, vector, matrix and higher-order representations (Section 1.3). Storing representations in high-order tensors allows associations between more than two words to be captured, like those found between word-link-word associations used to underpin a recent *distributional memory model* (DMM) [Baroni and Lenci, 2010].

The DMM was argued to provide a more flexible SSM (or distributional semantic model), by holding co-occurrence information about 3-tuples that can then be refined depending on the task being undertaken. When compared to strong benchmark models the DMM demonstrated effective performance on a wide range of tasks, including modeling word similarity judgments, discovering synonyms, concept categorization, solving analogy problems and classifying relations between word pairs. The DMM's reliance on linguistic resources places it out of the scope of this work. However, the use of tensor representations by the DMM to achieve effective performance on tasks, like solving analogy problems, and its breadth of application adds weight to the reported belief that the next generation of SSMs may require higher-order tensor representations to allow robust performance on a wider array of semantic tasks [Turney and Pantel, 2010].

Even though the DMM model is shown to provide robust performance across a wide range of semantic tasks, which is argued to be due to the use of the third-order tensor representations, there does not appear to be any evaluation of higher-order representations to confirm tuples are optimal. Possibly higher-order representations may lead to even further improvements. The greatest drawback of using higher-order tensor representations stems from the size of the representations [Beylkin and Mohlenkamp, 2002, 2005]. Within the DMM's framework the size of the model was constrained by restricting the tensor representation to *word-link-word* 3-tuples for a small vocabulary of 30,693 lemmas based on frequency cut-offs. Existing tensor models that have proposed the use of even higher-orders, have often only been presented theoretically.

A purely theoretical approach to using outer products of vectors to create higher-order tensor

representations was outlined in the matrix model of memory proposed by Humphreys et al. [1989]. However, this work did not address the practical considerations of implementing such a model and no empirical evaluation appears to have been reported.

BEAGLE's binding process (Section 2.6.1) compresses the outer products, and hence avoids what would otherwise have been the creation of higher order tensor representations, while still capturing some information about n-grams. However, the compression used by BEAGLE does not discriminate between informative co-occurrences and uninformative co-occurrences and hence considerable noise exists within the model and is likely to limit its effectiveness on a broad range of tasks.

The RI model (see Section 2.5.5) has also been used to capture co-occurrence information of n-tuples using outer products of the environment vectors [Sandin et al., 2011]. As RI is a fixed dimension approach this reduces the size of the tensors being produced. However, the tensor representations are still very sparse and grow considerably as the size of the n-grams are increased. Therefore, Sandin et al. [2011] developed a novel storage technique based on storing only the index positions of the non-zero elements, so as to reduce the storage complexity of the model. Given RI only uses a small number of non-zero elements in the environment vectors (usually less than ten), the resulting representations would be very sparse. It is worth noting that this method, or similar techniques for reducing the memory footprint of high-order tensor representations, could not be applied for fixed dimension models, like BEAGLE, as they use dense environment vectors.

However, the storage technique outlined by Sandin et al. [2011] provides a valuable insight into one approach for efficiently storing sparse, high-order tensor representations. A number of variants of this RI based model were evaluated on a benchmark synonym judgement test with varying effectiveness, none of which appeared to show improved effectiveness when compared to those report by SSMs using vector representations[3].

2.8 Summary

The evolution of SSMs, through the advent of dimension reduction techniques, order-encoding and higher-order tensor representations, has seen the robustness and range of applications they can be applied to increase. However, there does not appear to be any existing SSMs that use

[3]Listed at http://aclweb.org/aclwiki

all three evolutionary features. The following chapter outlines the development of an efficient, fixed dimension, order-encoding SSM that uses tensor representations. This next generation SSM is then used to underpin a formal model of word meaning based on the theories of structural linguistics.

Chapter 3

The Tensor Encoding (TE) Model

Ferdinand de Saussure (1916) argued that the differential view of meaning arose from two types of relationships that exist between linguistic concepts. These linguistic relationships, known as (i) syntagmatic and (ii) paradigmatic associations, underpin structural linguistic theory and have been argued to provide a relatively clean linguistic framework, free of psychology, sociology and anthropology [Holland, 1992].

These two types of associations have been used to motivate models of word meaning in the past [Dennis and Harrington, 2001], including those based on SSM technology [Sahlgren et al., 2008, Schütze and Pedersen, 1993]. The *tensor encoding* (TE) model, developed in this thesis, differs significantly from these in three main ways:

1. most importantly; the TE model presents a formal framework in which to model and combine explicit measures of syntagmatic and paradigmatic information; this has not been done before,

2. the TE model includes the use of 3 key advances in SSM technology, namely (i) the storage of representations within fixed dimension vectors, (ii) the encoding of word order; and (iii) the use of tensor representations.

3. The TE model does not rely on external linguistic resources, such as POS taggers, parsers or hand crafted knowledge sources.

The inspiration and design choices made during the development of this model are outlined in the following sections of this chapter. The development of the TE model can be broken down into two sections based on the two mathematical formalisms that facilitate (i) the construction

of its tensor representations, and (ii) the combination of two measures that explicitly model syntagmatic and paradigmatic associations.

3.1 Constructing Tensor Representations

3.1.1 The Binding Process

Inspired by BEAGLE (see Section 2.6.1), the TE model encodes word order and stores the representations within fixed dimension vectors. However, the encoding process does not compress the Kronecker products, and so tensor representations are formed during the binding process. To understand how this can be achieved without massive computational overhead a theoretical perspective needs to be taken.

Like BEAGLE (Section 2.6.1) and PRI (Section 2.6.2), within the TE model, the dimensionality of the vectors used to store the representations is fixed. However, within BEAGLE and PRI, the dimensionality of the environment vectors used to build the representations are the same as those of the final stored representations. Unlike BEAGLE and PRI, the TE model uses unary environment vectors to build the representations. This ensures an orthonormal basis is formed and the random variation in results experienced within BEAGLE and RI are removed. How is this achieved without large computational costs?

When these environment vectors are bound using Kronecker products, sparse, very large tensor representations are formed. In Section 3.1.4 it will be shown that these representations can be stored in relatively small fixed dimension storage vectors by using a novel dynamic compression technique. Experiments carried out on a number of semantic tasks in Chapter 4 demonstrate that this compression does not incur efficiency costs, and may even be responsible for increases in efficiency over other SSMs, including BEAGLE and PRI.

The representations formed within the TE model effectively contain both context and structural information within a single representation, known as the *memory tensors*. How this is achieved without the computational costs or random variation in results associated with BEAGLE is best illustrated by considering the binding process for the following example sentence, "*a dog bit the mailman*", where *a* and *the* are considered to be stop words (noisy, low information terms that are ignored) and hence will not be included in the vocabulary. The

resulting vocabulary terms and environment vectors become:

Term Id	Term	Environment vector
1	dog	$\boldsymbol{e}_{\text{dog}} = (1 \quad 0 \quad 0)^T$
2	bit	$\boldsymbol{e}_{\text{bit}} = (0 \quad 1 \quad 0)^T$
3	mailman	$\boldsymbol{e}_{\text{mailman}} = (0 \quad 0 \quad 1)^T$

The *memory tensor* for each term in the vocabulary is constructed by summing the resulting Kronecker products of the environment vectors within a sliding context window over the text. The number of environment vectors bound using Kronecker products impacts the order of the memory tensors. To illustrate, consider the binding process that would capture word order and co-occurrence information of 2-tuples within second-order tensor (matrix) representations:

$$\boldsymbol{M}_w = \sum_{k \in \{C | k \prec w\}} \boldsymbol{e}_k \otimes \boldsymbol{e}_w^T + \sum_{k \in \{C | w \prec k\}} \boldsymbol{e}_w \otimes \boldsymbol{e}_k^T, \tag{3.1}$$

where C is a totally ordered set of terms created by the sliding context window, containing two order relations $k \prec w$ and $w \prec k$, where $w \in C$ is the target term, $k \in C$ is a non-stop word found within the context window, and $k \prec w$ indicates that term k appears before term w in C. Note, stop words are not bound, but are included when determining the context window boundaries. This is done, as many existing corpus-based methods, including HAL (refer to Section 2.5.3) include stop words in determining the window boundaries. Choosing to ignore them during the binding process stems from the fact that by definition a stop word is ignored when it comes to computing similarities between representations, hence ignoring them in the binding process saves computational overhead.

The number of words contained within a context window is often ambiguously defined. In some studies it is referred to as window *size* [Lund and Burgess, 1996], window *length* [Bruza and Song, 2002], and others simply $\pm x$ [Rapp, 2002], where x is the number of terms either side of the target term, and the $\pm$ sign indicates before and after the target term. However, to clarify the important role that distance between the target term w and any other term in the window plays, the size of the context window will be referred to as *radius*.

The radius is the number of words between the target term, which is always at the centre of the window, and the edge of the window in either direction. To illustrate, consider the example

sentence given earlier when a context window of radius one is centred around the target word *bit* (shown in square brackets):

Example Context Window: $A_s \quad \overbrace{dog \quad [bit] \quad the_s} \quad mailman$

For the second-order TE model, the resulting memory tensors are effectively 3×3 matrices, having elements equal to the co-occurrence frequency of the 2-tuples formed by the target term and the terms found within the context window. To illustrate, consider the memory matrices created for the vocabulary terms using a sliding context window of radius 2 and with the target term shown in square brackets.

Binding Step 1: $\overbrace{A_s \quad [dog] \quad bit \quad the_s} \quad mailman$

$$\boldsymbol{M}_{dog} = \boldsymbol{e}_{dog} \otimes \boldsymbol{e}_{bit}^T = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (0 \quad 1 \quad 0) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Binding Step 2: $\overbrace{A_s \quad dog \quad [bit] \quad the_s \quad mailman}$

$$\begin{aligned} \boldsymbol{M}_{bit} &= \boldsymbol{e}_{dog} \otimes \boldsymbol{e}_{bit}^T + \boldsymbol{e}_{bit} \otimes \boldsymbol{e}_{mailman}^T \\ &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (0\,1\,0) + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} (0\,0\,1) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned}$$

Binding Step 3: $A_s \quad dog \quad \overbrace{bit \quad the_s \quad [mailman]}$

$$\boldsymbol{M}_{mailman} = \boldsymbol{e}_{bit} \otimes \boldsymbol{e}_{mailman}^T = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} (0 \quad 0 \quad 1) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The resulting pattern is that all non-zero elements are situated on the row or column corresponding to the target term's term-id. If this vocabulary building process was performed over the entire corpus the general form of a *memory matrix* would be:

$$\boldsymbol{M}_w = \begin{pmatrix} 0, & \ldots, 0, & f_{1w}, & 0, & \ldots, 0 \\ & & \ldots & & \\ 0, & \ldots, 0, & f_{(w-1)w}, & 0, & \ldots, 0 \\ f_{w1}, \ldots, f_{w(w-1)}, & & f_{ww}, & f_{w(w+1)}, \ldots, f_{wn} & \\ 0, & \ldots, 0, & f_{(w+1)w}, & 0, & \ldots, 0 \\ & & \ldots & & \\ 0, & \ldots, 0, & f_{nw}, & 0, & \ldots, 0 \end{pmatrix}, \tag{3.2}$$

where f_{iw} is the value in row i column w of the matrix representing the ordered co-occurrence frequencies of term i before term w, f_{wj} is the value in row w column j of the matrix that represents the ordered co-occurrence of term j after term w, and n is the number of unique terms in the vocabulary.

Another advantage of the TE model's binding process over current fixed dimension approaches, like BEAGLE and RI, is the existence of explicit co-occurrence frequencies within the geometric representations, as seen in Equation (3.2). Being able to access explicit co-occurrence information allows information theoretic measures (based in probability theory) to be used in addition to geometric measures when performing similarity judgements between words in the model. The merging of probabilistic and geometric models is argued to be a developing feature in the evolution of SSMs [Turney and Pantel, 2010]. Recent work has also demonstrated the effectiveness of merging geometric and probabilistic measures on a number of semantic categorization tasks [Bullinaria and Levy, 2007].

This use of either geometric or probabilistic measures can create some confusion around the classification of the model as an SSM. In these cases a more general class, known as *distributional semantic models* (DSM) may be more accurate, and has been used in the past to encompass models based on the distributional hypothesis [Baroni and Lenci, 2010] (Section 1.1). Even if having access to both types of mathematical measures places the TE model in this broader class, it is important to acknowledge that the formal binding process used to build representations within the TE model stores them within a geometric space.

3.1.2 Using High-order Tensor Representations

The TE model can be extended to capture higher-order co-occurrence information of n-tuples. An n-tuple is formed by combining n words seen in order within the context window. Within BEAGLE, n-grams are formed by binding n ordered words within the context window. The difference between an n-tuple and an n-gram is that the words in a tuple do not need to appear directly beside each other in language, whereas the words in n-grams do. However, in both cases the order of the words is maintained. Within the TE model the n^{th}-order binding process captures explicit co-occurrence frequencies of words found in n-tuples.

To illustrate, consider the example sentence: "*The dog bit the mailman viciously*", which has the following vocabulary terms and environment vectors:

Term Id	Term	Environment vector
1	dog	$e_{dog} = (1 \quad 0 \quad 0 \quad 0)^T$
2	bit	$e_{bit} = (0 \quad 1 \quad 0 \quad 0)^T$
3	mailman	$e_{mailman} = (0 \quad 0 \quad 1 \quad 0)^T$
4	viciously	$e_{viciously} = (0 \quad 0 \quad 0 \quad 1)^T$

For the third-order TE model, the binding operation can be defined as:

$$\begin{aligned} M_w = & \sum_{k_1,k_2 \in \{C | k_1 \prec w, k_2 \prec w, k_1 \prec k_2\}} e_w \otimes (e_{k_1} \otimes e_{k_2}^T) + \\ & \sum_{k_1,k_2 \in \{C | k_1 \prec w, w \prec k_2, k_1 \prec k_2\}} e_{k_2} \otimes (e_{k_1} \otimes e_w^T) + \\ & \sum_{k_1,k_2 \in \{C | w \prec k_1, w \prec k_2, k_1 \prec k_2\}} e_{k_2} \otimes (e_w \otimes e_{k_1}^T), \end{aligned} \tag{3.3}$$

where $k_1, k_1 \in C$ denote vocabulary terms within set of context window terms C, w is the target term and $(C, \prec)$ is an irreflexive, anti-symmetric and transitive binary relation where $x \prec y$ denotes word x is before word y. An important constraint on this binding process is that there are always three terms used in forming the Kronecker products. This is to ensure the same order tensors are produced as required for addition of the resulting tensors.

Consider the resulting memory tensor for the term "*bit*", using the example sentence provided earlier and a sliding context window radius of 3:

Binding Step for focus term *bit*:

$$\overbrace{The_s \quad dog \quad [bit] \quad the_s \quad mailman \quad viciously}$$

$$\begin{aligned}
M_{bit} &= e_{mailman} \otimes (e_{dog} \otimes e_{bit}^T) + e_{viciously} \otimes (e_{bit} \otimes e_{mailman}^T) \\
&= \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \left[\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}^T \right] + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \left[\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}^T \right] \\
&= \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
M_{bit} &= \left(\begin{array}{c} 0_{4\times4} \\ \hline 0_{4\times4} \\ \hline \begin{matrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix} \\ \hline 0_{4\times4} \end{array} \right) + \left(\begin{array}{c} 0_{4\times4} \\ \hline 0_{4\times4} \\ \hline \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix} \end{array} \right) = \left(\begin{array}{c} 0_{4\times4} \\ \hline 0_{4\times4} \\ \hline \begin{matrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix} \\ \hline \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix} \end{array} \right)
\end{aligned}$$

It can be seen from the resulting memory tensor, M_{bit} that the Kronecker products of terms 1 (dog), 2 (bit) and 3 (mailman), and terms 2 (bit), 3 (mailman) and 4 (viciously) store the frequency of the 3-tuples (1,2,3 - *dog bit mailman*) and (2,3,4 - *bit mailman viciously*), respectively. This is indicated by non-zero elements existing in row **1**, column **2** of matrix **3**, and in row **2**, column **3** of matrix **4** of the memory tensors. Within each of these ordered 3-tuples there exists information about ordered 2-tuples, including those between terms (1,2), (2,3); (2,3) and (3,4). This means that a third-order model also includes all of the co-occurrence information found in the second-order model. Storing co-occurrence information of 3-tuples will allow the TE model to have access to a flexible array of associational information, as used in the Distributional Memory Model (DMM) [Baroni and Lenci, 2010] (Section 2.7), and in a similar way can be argued to make the TE model applicable across a wider array of tasks.

Due to the sparseness of the memory tensors created by the binding process and the existence of element values that represent the explicit co-occurrence frequencies, the building and efficient storage of these memory tensors can be achieved using compression ideas similar to that of Sandin et al. [2011]. Before presenting the development of our novel memory tensor compression technique, an extension to the TE model's binding process that allows richer proximity information to be incorporated within the representations is provided.

3.1.3 Capturing Richer Proximity Information

Using a sliding context window ensures some proximity information between terms in the window is *implicitly* captured. However, like the HAL model (Section 2.5.3), the TE model can capture stronger proximity information by weighting the strength of co-occurrences inversely proportional to the distance between the target term and the other terms in the context window. The choice of proximity weighting function will have implications for computational complexity and likely the effectiveness of the model.

To illustrate the impact on computational complexity, consider the linear proximity weighting function used by the HAL model (Section 2.5.3), which can be implemented within the TE model's binding process as:

$$f(d_k) = R - d_k + 1, \tag{3.4}$$

where R is the radius of the context window (see Section 3.1.1) and d_k is the distance between the target term and term being considered. Using this function the resulting elements within

the memory tensor representations would be integers ($\mathbb{I}$). Ensuring all elements remain integers provides a significant computational complexity benefit over the use of proximity weighting functions that result in real ($\mathbb{R}$) elements within the representations, such as those created by a rational function like:

$$f(d_k) = \frac{1}{d_k}. \tag{3.5}$$

This is due to the storage requirements of integers being much less than reals within computer memory, as well as the savings in processing time associated with integer computations compared to those involving real numbers.

Incorporating the linear proximity weighting function in Equation (3.4) into the binding process in Equation (3.1) gives:

$$\boldsymbol{M}_w = \sum_{k \in \{C | k \prec w\}} (R - d_k + 1).\boldsymbol{e}_k \otimes \boldsymbol{e}_w^T + \sum_{k \in \{C | w \prec k\}} (R - d_k + 1).\boldsymbol{e}_w \otimes \boldsymbol{e}_k^T, \tag{3.6}$$

where R is the radius of the sliding context window, and d_k is the distance between term k and target term w found within the set of terms in the sliding context window C. The distance between terms is measured by the number of gaps between terms, or jumps to get from term w to term k. To demonstrate, consider our previous example sentence, noting *bit* and *mailman* are 2 terms apart in the sentence (as stop words are included when calculating distance within the context window):

Binding Step (with proximity weighting): $\overbrace{A_s \quad dog \quad [bit] \quad the_s \quad mailman}$

$$\begin{aligned} \boldsymbol{M}_{bit} &= 2 \times \boldsymbol{e}_{dog} \otimes \boldsymbol{e}_{bit}^T + \boldsymbol{e}_{bit} \otimes \boldsymbol{e}_{mailman}^T \\ &= 2 \times \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (0\ 1\ 0) + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} (0\ 0\ 1) = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{3.7}$$

The linear proximity weighting function used in the TE model's binding process can also be replaced by other types of functions, including the rational function in Equation (3.5), giving:

$$\boldsymbol{M}_w = \sum_{k \in \{C | k \prec w\}} \frac{1}{d_k} \boldsymbol{e}_k \otimes \boldsymbol{e}_w^T + \sum_{k \in \{C | w \prec k\}} \frac{1}{d_k} \boldsymbol{e}_w \otimes \boldsymbol{e}_k^T. \tag{3.8}$$

To gain an understanding of the impact that the choice of proximity weighting function has on the effectiveness of the TE model, the performance of the TE model on a semantic distance

task, using no proximity weighting (TE_noScale), linear proximity weighting (TE_linear, Equation (3.4)) and a rational linear weighting function (TE_1overX, Equation (3.5)) is shown in Figure 3.1. The experimental setup for this semantic distance task is outlined in Section 4.3.

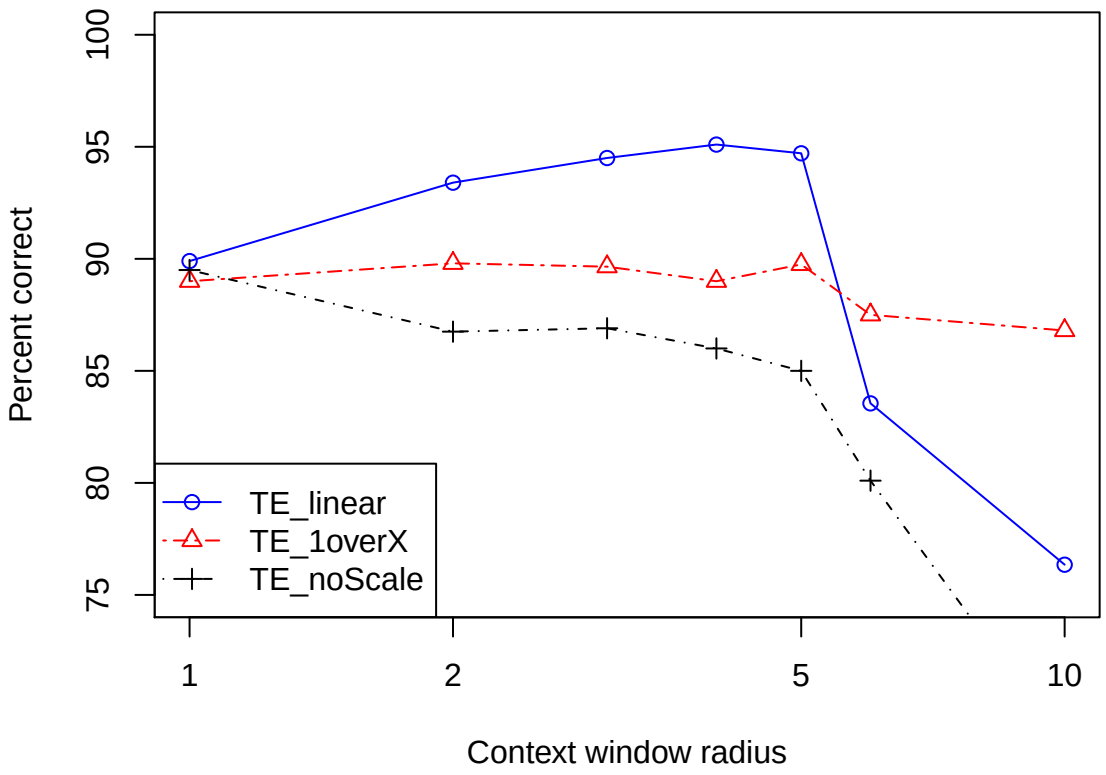

Figure 3.1: Performance of the second-order TE model on a semantic distance task for various proximity weighting functions across various context window radii.

Figure 3.1 suggests that the linear proximity weighting function (Equation (3.4)) provides superior task effectiveness for smaller context window lengths. However, the effectiveness achieved by the TE model using the rational weighting function (Equation (3.5)) appears less sensitive to changes in radius of the context window (R) used to build up the representations. Both, the linear and rational proximity weighting functions provide superior task effectiveness when compared to a TE model that does not implement any proximity weighting (TE_noScale in Figure 3.1).

Before leaving this discussion on proximity weighting it is worth illustrating how this would be incorporated into higher-order binding operations. For example, the third-order binding process in Equation (3.3) could be updated to include the linear proximity weighting function

as follows:

$$M_w = \sum_{k_1,k_2 \in \{C | k_1 \prec w, k_2 \prec w, k_1 \prec k_2\}} e_w \otimes ((R - d_{k_1} + 1).e_{k_1} \otimes (R - d_{k_2} + 1).e_{k_2}^T) \quad + \\ \sum_{k_1,k_2 \in \{C | k_1 \prec w, w \prec k_2, k_1 \prec k_2\}} (R - d_{k_2} + 1).e_{k_2} \otimes ((R - d_{k_1} + 1).e_{k_1} \otimes e_w^T) \quad + \\ \sum_{k_1,k_2 \in \{C | w \prec k_1, w \prec k_2, k_1 \prec k_2\}} (R - d_{k_2} + 1).e_{k_2} \otimes (e_w \otimes (R - d_{k_1} + 1).e_{k_1}^T), \tag{3.9}$$

where d_{k_1} is the number of words between the target term w and term k_1, and d_{k_2} is the number of words between the target term w and term k_2.

Using the example sentence provided in Section 3.1.2 (*The dog bit them mailman viciously*) and the third-order binding process in Equation (3.9), the resulting memory tensor for M_{bit} becomes:

Binding Step for focus term *bit*:

$$\overbrace{The_s \quad dog \quad [bit] \quad the_s \quad mailman \quad viciously}$$

$$M_{bit} = 2e_{mailman} \otimes (3e_{dog} \otimes e_{bit}^T) + e_{viciously} \otimes (e_{bit} \otimes 2e_{mailman}^T)$$

$$= 2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \otimes \left[3 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}^T \right] + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \otimes \left[\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} 2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}^T \right]$$

$$= 2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$M_{bit} = \left(\begin{array}{cccc} \multicolumn{4}{c}{0_{4\times4}} \\ \hline \multicolumn{4}{c}{0_{4\times4}} \\ \hline 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline \multicolumn{4}{c}{0_{4\times4}} \end{array}\right) + \left(\begin{array}{cccc} \multicolumn{4}{c}{0_{4\times4}} \\ \hline \multicolumn{4}{c}{0_{4\times4}} \\ \hline \multicolumn{4}{c}{0_{4\times4}} \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) = \left(\begin{array}{cccc} \multicolumn{4}{c}{0_{4\times4}} \\ \hline \multicolumn{4}{c}{0_{4\times4}} \\ \hline 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \tag{3.10}$$

The magnitude of the element values in the third-order tensor representation, M_{bit} provide an indication of the combined proximity of the terms bound to the target term. With terms 1 (dog) and 3 (mailman) being overall in closer proximity to term 2 (bit) than terms 3 and 4 (viscously) within the sentence. This is indicated by the first tuple (1-2-3) having an element value of six and the second tuple (2-3-4) having a value of two[1].

3.1.4 Efficient Tensor Computations

One of the most significant drawbacks associated with models using high-order tensor representations (when compared to vector representations) stems from the computational complexities commonly associated with them [Beylkin and Mohlenkamp, 2005]. This is due to the space required to store the representations and the processing time associated with performing computations involving these representations.

[1]It is important to note that in the design of the TE model, stop words are counted when calculating d_{k_i}, the distance between term k_i and the target term w in Equation (3.9), refer to Section 3.1.1.

To this point, the development of the TE model's semantic space has resulted in word order and co-occurrence information being stored within very sparse, high dimensional tensors, whose order increases as the number of terms involved in the binding operation increases. Typically these high-order representations have made tensor based SSMs intractable. However, a novel compression technique will now be presented, that takes advantage of the sparseness in the TE model representations (Section 3.1.1), and effectively allows them to be stored in fixed dimension vectors.

To demonstrate, consider the memory matrix for *bit* in the proximity scaled example of Equation (3.7), reproduced here for clarity:

$$\boldsymbol{M}_{bit} = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

can be stored in the following fixed dimension *storage vector* (SV):

$$\text{SV}_{\text{bit}} = [(-1 \quad 2) \quad (3 \quad 1)], \tag{3.11}$$

where parenthesis have been added to illustrate implicit grouping of $(T \quad CF)$ pairs, where T is the term-id of the co-occurring term and CF is the cumulative, proximity-scaled, co-occurrence frequency of T with w. The sign of T (term-id) indicates the word order of T with w. Knowing that a context window of radius 2 was used, the storage vector above indicates that the word *dog* (term T $= 1$) appeared directly before (as indicated by the negative sign) the word *bit* (hence, CF $= 2$), and the word mailman (term T $= 3$) occurred two words after *bit* (hence, CF $= 1$). The information in this vector can be used to reconstruct the memory matrix using the following process:

1. If the term Id (T) in the $(T \quad CF)$ pair is positive, the CF value is located at row w, column T in the memory tensor. Otherwise, the CF value is located at row T, column w.

At an implementation level, the construction of the second-order representations can be efficiently achieved using fixed dimension storage vectors and the following process:

1. For each co-occurrence with target term w, search the $(T \quad CF)$ pairs in the storage vector (SV_w) for a matching T value and its sign to ensure it occurs in the same word order with w.

2. If a match is found then, the CF element of the pair is increased by the scaled, co-occurrence frequency of w with T within the current context window. End process.

3. If no match is found then, check if the storage vector is full

4. If the storage vector is full then, the first low information $(T \quad CF)$ pair in the storage vector should be removed and the new $(T \quad CF)$ pair added to the end of the storage vector.

5. If the vector is not full then add the new $(T \quad CF)$ pair to the end of the storage vector.

The removal of the first, low information $(T \quad CF)$ pair in the storage vector, when the vector is full, applies a form of compression to the model, which I call *tensor memory compression* (TMC). An efficient measure of low information can be expressed as:

$$I_{wT}(t) = \frac{CF_{wT}(t)}{F_w(t)}, \tag{3.12}$$

where $CF_{wT}(t)$ is the ordered co-occurrence frequency of term T and w at time t, and $F_w(t)$ is the collection frequency of the target term w at time t. If no $(T \quad CF)$ pairs within the storage vector have an $I_{wT}(t)$ estimate less than the chosen *co-occurrence frequency cut-off*, then the current interaction is discarded. This process is aimed at increasing the information content of the storage vectors.

Even for applications where the vocabulary is small and the context window radius is small, there will be a number of low information terms in the corpus. These terms either co-occur with many terms often (e.g., *the*), and hence will fill the storage vectors quickly, or appear so infrequently that estimates using these values will be statistically unreliable. SSMs often use techniques such as frequency cut-offs or the use of a stoplist (i.e., a list of high frequency terms that are to be omitted from the vocabulary building process) to reduce the impact of low information co-occurrences associated with high frequency or very low frequency terms [Bullinaria and Levy, 2007, Rohde et al., 2006].

The TMC technique, when used in combination with a stoplist, allows the TE model to handle these low information terms within fixed dimension storage vectors. The main difference between the compression within the TE model and current fixed dimension approaches, like BEAGLE, is that the TMC technique uses information sensitive compression. It is hypothesised that this information sensitive compression, and hence the TMC process, will be an important

factor in allowing the TE model to achieve superior task effectiveness. This novel compression technique is considered a significant contribution of this research to the field and since the process is not specific to the TE model, it could be exploited by other corpus-based approaches, such as HAL-based models.

Further support for compressing lexical representations comes from findings that within the human brain some form of information compression is used [Machens, 2012].

Tensor Memory Compression

To explain the TMC process further, recall that as the storage vector is filled from the first empty position, a temporal property is created within the storage vector, with the first element positions having been filled before the following and so on. By searching from the first position of the storage vector and finding the first (T CF) pair that is below the co-occurrence frequency cut-off, this pair will be the oldest pair below the cut-off that has co-occurred with the target term w. If the storage vector is full, the pair below the co-occurrence frequency cut-off can be removed and the new (T CF) pair added to the end of the storage vector. Adding this new interaction to the end of the list reduces the likelihood that this new memory will be replaced next.

According to a recent literature survey it appears that this type of temporal property created by the TMC technique does not exist in existing corpus-based SSMs. For example, within BEAGLE, HAL, LSA and RI, a co-occurrence with a target term w that is processed early in the vocabulary building process has the same impact on the context vector of w as the very last co-occurrence for target term w. This may not be the case for a memory tensor within the TE model, especially considering the early co-occurrence may have been long discarded from the representation. This temporal property created by the TMC process, means that the TE model is better able to encapsulate the dynamic properties of meaning described by structural linguists, like Saussure and Peirce (refer to Section 2.2).

It is worth noting that by modifying the calculation of low-information, many variants of the TMC technique could be developed. For example, one may wish to create a different temporal effect by weighting the co-occurrence frequency cut-off by the position of the (T CF) pair within the storage vector. An example function to achieve this would be:

$$I_{\text{TMC}}(t) = \frac{CF(t)}{F_w(t)(D_{\text{SV}} - \frac{p}{2})}, \tag{3.13}$$

where D_{SV} is the dimensionality of the storage vector, and p is the position in the storage vector, with zero being the start position, and $(\frac{D_{\text{SV}}}{2} - 1)$ being the end position of the vector.

Research into effective policies for measuring informativeness have been developed for applications, including entity detections [Rennie and Jaakkola, 2005] and index pruning [Carmel et al., 2001]. However, as the TMC technique proposed in Equation (3.12) allows us to efficiently work with higher-order tensor representations, further investigations into various TMC algorithms is left for future work.

For the experiments carried out in this research the value of co-occurrence frequency cut-off is fixed, in effect removing it from the list of free model parameters. This improves variable control and helps improve confidence with which changes in task effectiveness can be attributed. To briefly demonstrate the impact that fixing the co-occurrence frequency cut-off value may have on task effectiveness, the performance of the TE model on a synonym judgement task is shown in Figure 3.2. The experimental set up of this task is outlined in Section 4.2.

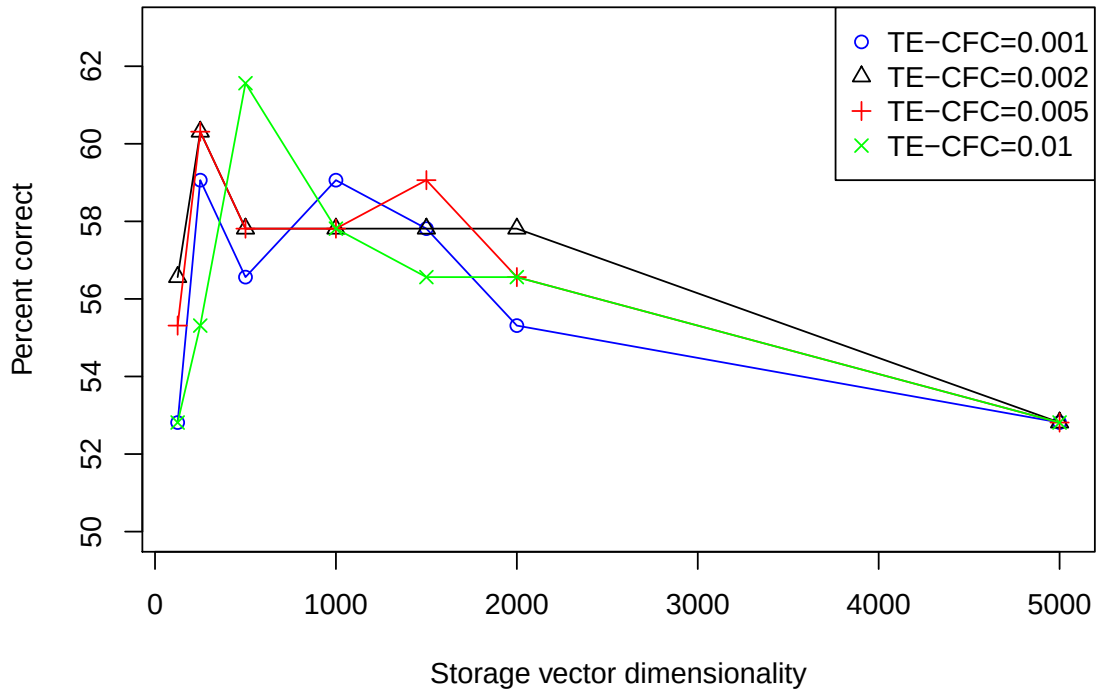

Figure 3.2: Performance of the TE model on a synonym judgement test using a small training corpus for various storage vector dimensions and co-occurrence frequency cut-off (CFC) values. The evaluation was carried out using a context window radius of 1 and $\gamma = 0.4$.

This graph illustrates how the TMC approach allows the TE model to achieve superior

effectiveness using lower dimensional storage vectors, with the best (or equal best) effectiveness being produced for storage vectors of no more than 500 dimensions. The best performance was produced when using a co-occurrence frequency cut-off equal to 0.01, which meant that when a storage vector was full, the first term that co-occurred less than 1 in every 100 occurrences of the target term would be discarded.

Based on the similar robustness achieved by TE model for all co-occurrence frequency cut-off values and storage vectors less than 2,000 dimensions, a co-occurrence frequency cut-off equal to 0.002 is argued to be a reasonable default value for use in the experiments performed in this research.

Efficiently Storing higher-order Tensors

Before moving on to quantify the efficiency of the TE model, it is worth illustrating how higher-order TE model representations are stored at an implementation level. The storage vector for the third-order representation of *bit*, shown in Equation (3.10), reproduced here:

$$M_{bit} = \left(\begin{array}{c} 0_{4\times4} \\ \hline 0_{4\times4} \\ \hline \begin{array}{cccc} 0 & 6 & 0 & 0 \end{array} \\ \begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \\ \hline \begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccc} 0 & 0 & 2 & 0 \end{array} \\ \begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \\ \begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \end{array}\right)$$

can be given as:

$$\text{SV}_{\text{bit}} = [(-1 \quad 3 \quad 6) \quad (3 \quad 4 \quad 2)] \tag{3.14}$$

where parenthesis have been added to illustrate implicit grouping of (T1 T2 CF) triples. The information in this vector can be used to reconstruct the memory tensor in Equation (3.10) using the following process:

1. If T1 and T2 are positive, then the CF value is located at row w, column T1 and matrix T2 within the third-order tensor representation. If T1 is negative and T2 is positive, then the CF value is located at row T1, column w and matrix T2. If T1 and T2 are negative, then the CF value is located at row T1, column T2 and matrix w[2].

Computational Complexity of Building the Semantic Space

The storage and time complexity of models working with tensors is often polynomial with the order of the tensors, i.e., for vectors $O(n)$, matrices $O(n \times n)$, third-order tensors $O(n \times n \times n)$. To demonstrate the efficiency gains associated with using the TMC technique within the storage vectors of the TE model, a storage and time complexity analysis for the second-order TE model (i.e., storing memory matrices) is undertaken.

Storage Complexity: For the second-order TE model the memory matrices are implemented as low dimensional storage vectors with dimensionality D_{SV}. This means that for a vocabulary of $|V| = n$ terms, the storage complexity of the second-order TE model is $M(n) = n.D_{SV}$. The graph in Figure 3.2 shows that optimal performance can be achieved for storage vectors with dimensions as low as 200, and often up to 500. This means the storage complexity of the TE model is $M(n) = 200.D_{SV}$

Given BEAGLE and PRI have been reported to use context and order vectors in excess of 2,000 dimensions to achieve their best performance on this same task and corpus [Jones and Mewhort, 2007, Sahlgren et al., 2008] the use of the TMC technique within the TE model appears to provide a storage complexity advantage. It is worth noting that dimension reduction (i.e., SVD) has been applied to RI (Section 2.5.5) in the past, in an attempt to further reduce the size of the context vectors [Lin and Gunopulos, 2003]. However, this reduction in storage

[2] As T1 always precedes T2 in the binding operation is not possible for T2 to be negative at the same time as T1 is positive.

complexity did not provide consistent improvements in time complexity or task accuracy. The storage advantage of TMC is also seen when compared to the memory footprint of HAL-based models. These approaches often use storage vectors in the order of tens of thousands, if not hundreds of thousands [Bullinaria and Levy, 2007]. Even though LSA uses context vectors with fewer dimensions, often 200-300 [Landauer and Dumais, 1997], the full co-occurrence matrix is still required to be built before SVD can be applied, and hence LSA has a total memory footprint larger than most fixed dimension approaches.

Another interesting point of contrast is seen when considering the amount of redundant information in the TE model representations. Examination of the general form of second-order TE memory tensors in Equation (3.2) suggests that the co-occurrence information between two terms w_1 and w_2, exists in the memory matrices of both w_1 and w_2. Given (i) the HAL matrix contains no redundant information (ii) the TE model has a much smaller storage footprint than HAL-based models, and (iii) experiments in Chapter 4 show the TE model achieves superior effectiveness over a strong HAL-based approach on a number of semantic tasks; it is important to understand how the TE model gains this efficiency advantage over other SSMs without losing effectiveness.

One point that may explain such a contrast is that the redundancy within the TE model's memory tensors is only theoretical, and may be absent in practice due to the TMC process. However, using TMC also means that if redundant information exists across the memory tensors it is likely to be informative. By design, the redundancy in the memory tensors removes a layer of computational complexity that would be required to track and access relevant information within the representations when performing similarity calculations. For example, if trying to identify the terms most likely to co-occur with term w_1 in the vocabulary, only the terms found in the storage vector of w_1 need to be considered, as compared to the full search of the vocabulary used by RI or BEAGLE. A full search involves comparing the target term's representation to all other representations in the vocabulary. By avoiding this full search, the time complexity of the TE model is improved when compared to current SSMs, as will be highlighted in Section 3.2.2.

In Chapter 4, it will be shown that the TMC technique allows the TE model to achieve superior effectiveness on a range of semantic tasks when compared to a number of strong benchmark models, including a state-of-the-art HAL-based model. Therefore, it could be proposed that the TMC technique could also serve as an efficient method for dimension reduction within HAL-based models (Section 2.5.3).

Time Complexity: To determine whether the binding process of the TE model is more efficient than the BEAGLE and PRI models, a time complexity analysis is also required. The time complexity of the TE model's vocabulary building operation is determined by considering the worst case, which occurs when the storage vectors are full and a replacement operation is performed. In this case, the basic operation of the binding process becomes a full search of the (T CF) list, giving: $T_{TE}(n) = O(\frac{D_{SV}}{2})$, where D_{SV} is the storage vector dimensionality.

For the task of synonym judgement, used to produce Figure 3.2, the optimal effectiveness of the TE model using a co-occurrence frequency cut-off equal to 0.002 is achieved when $D_{SV} = 250$, therefore, $T_{TE}(n) = 125$. To gauge whether this is efficient, lets compare it to the binding process used by the permuted random indexing (PRI) model (Section 2.6.2), which is a fixed dimension approach that encodes word order and has been shown to be more efficient than BEAGLE.

PRI's binding process involves the summing of an environment vector with a context vector, and a permuted environment vector with an order vector. Assuming the dimensionality of the vectors are D_{PRI}, the time complexity of the PRI model would be $T_{PRI}(n) = O(2.D_{PRI})$. Optimal performance of PRI on a synonym judgement task was achieved when $D_{PRI} = 30,000$. Therefore, $T_{PRI}(n) \geq 64,000$. Hence, for the task of synonym judgement, the second-order TE model appears to build its semantic space $\frac{64,000}{125} = 512$ times more efficiently than the PRI approach. Further comparison of computational complexities between SSMs will be provided when considering the costs associated with computing similarity between words within the TE model (Section 3.2.2).

Summary

The computational complexity analysis provided here suggests that the TE model has reduced storage complexity compared to current SSMs, and can build its semantic with much lower time complexity than fixed dimension approaches, like BEAGLE and RI. Models using dimension reduction techniques, like SVD, have been shown to be even more computationally expensive when it comes to building the semantic space due to the cost of performing SVD. To determine whether the TE model's efficiency advantage can be delivered without compromising effectiveness, the performance of the TE model will be compared against a number of strong benchmark SSMs on various semantic tasks in Chapter 4.

The building of the TE model's underlying vocabulary has involved a unique binding process, based on uncompressed Kronecker products, and a novel method for compressing the resulting tensors, known as the TMC technique. The following section outlines how structural linguistic theory can be used to motivate a novel approach to modelling and combining information about word associations stored in these representations.

3.2 Modelling Word Meaning

Within SSMs, once the semantic space has been constructed, information regarding associations between words is often accessed using a single measure of semantic similarity. The measures chosen are often dependent on the task, as they often require different types of information to achieve the task successfully. For example, the task of synonym judgement relies more heavily on information about paradigmatic associations between words. This can be seen by noting that the word *postman* would likely have similar close neighbours as the word *mailman* and hence would be able to replace it in the following sentence: *A dog bit the* ***mailman***, such that a paradigmatic association between *postman* and *mailman* can be said to exist. However, identifying measures that are effective across many tasks can be difficult.

A comprehensive analysis of the effectiveness of fourteen different similarity measures on four semantic tasks was carried out by Bullinaria and Levy [2007]. This work highlights the difficulty in selecting a single measure to use across all tasks, and the lack of any formal link back to a definition of linguistic meaning in making similarity judgements.

The approach developed in this work provides a formal measure of similarity in meaning based on structural linguistics. More specifically, the meaning of a concept is based on the syntagmatic and paradigmatic associations it exhibits, and that by comparing these associations between concepts one can determine the *similarity in meaning of one concept with another*. Within this research the concept under consideration is that of a *word*, however, much of the theory developed can be generalised to abstract concepts, even outside of linguistics.

3.2.1 A Formal Framework

To formally combine information about syntagmatic and paradigmatic associations between two words an undirected graph, specifically a *Markov random field* (MRF) can be used. A MRF is an

undirected graph that formally represents the dependencies between random variables [Koller and Friedman, 2009]. MRF's have been used to successfully underpin feature based models within the field of information retrieval [Metzler and Croft, 2007].

Using an MRF, a conditional probability estimate for observing a vocabulary word w given some priming word q can be produced. The proof begins by letting an undirected graph G contain nodes that represent random variables, and letting the edges define the independence semantics between the random variables. Within the graph, a random variable is independent of its non-neighbours given observed values of its neighbours.

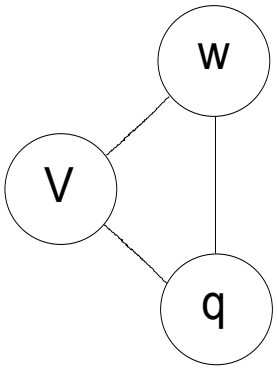

Figure 3.3: Example Markov random field for the TE model.

Figure 3.3 shows a graph G that consists of word node q, and word node w, and a vocabulary node V. Words q and w are constrained to exist within the vocabulary V. We parameterize the graph based on clique sets to provide more flexibility in encoding useful features over cliques in the graph. The joint distribution over the random variables in G is defined by:

$$P_{G,\Gamma}(q, w, V) = \frac{1}{Z_\Gamma} \prod_{c \in cl(G)} \varphi(c; \Gamma), \tag{3.15}$$

where $cl(G)$ is the set of cliques in G, each $\varphi(.; \Gamma)$ is a non-negative *potential function* over clique configurations parameterized by Γ, and $Z_\Gamma = \sum_{q,w} \prod_{c \in cl(G)} \varphi(c; \Gamma)$ normalizes the distribution. The joint distribution is uniquely defined by the graph G, potential functions φ and the parameter Γ. Given the logarithm of products is equal to the sum of logarithms, the simplified form of the joint distribution becomes:

$$\log P_{G,\Gamma}(q, w, V) = \frac{1}{Z_\Gamma} \sum_{c \in cl(G)} \log \varphi(c; \Gamma), \tag{3.16}$$

where the potential functions are commonly parameterized as:

$$\varphi(c;\Gamma) = \exp[\gamma_c f(c)], \tag{3.17}$$

with $f(c)$ being some real-valued *feature function* over clique values and γ_c is the weight given to that particular feature function. Substituting Equation (3.17) into Equation (3.16) gives:

$$\log P_{G,\Gamma}(q, w, V) = \frac{1}{Z_\Gamma} \sum_{c \in cl(G)} \gamma_c f(c). \tag{3.18}$$

After G is constructed, the conditional probability of observing word w given q can be expressed as:

$$P_{G,\Gamma}(w|q) = \frac{P_{G,\Gamma}(q, w, V)}{\sum_{w \in V} P_{G,\Gamma}(q, w, V)}, \tag{3.19}$$

where V is the universe of all possible vocabulary terms and w is a possible vocabulary term.

By using Equation (3.18) and Equation (3.19) with constant terms removed, a rank equivalent form for the conditional probability can be written as:

$$P_{G,\Gamma}(w|q) \propto \sum_{c \in cl(G)} \gamma_c f(c), \tag{3.20}$$

where a constraint of $\sum_{c \in cl(G)} \gamma_c = 1$ is applied for ease of training.

Model Parameterization

The conditional probability expressed in Equation (3.20), provides a formal method for combining feature functions, designed to extract various types of word associations, mapped via cliques in the graph. For the graph shown in Figure 3.3, a number of useful clique sets capturing dependencies are summarised in Table 3.1.

Since it is not our goal to find optimal feature functions, but to demonstrate the use of a Markov random field to formally combine feature functions that model syntagmatic and paradigmatic associations, we focus on evaluating estimates over the clique sets relevant to the syntagmatic and paradigmatic measures.

Assuming some syntagmatic feature $s_{\text{syn}}(q, w)$ and paradigmatic feature $s_{\text{par}}(q, w)$ are chosen and using the T_{syn} and T_{par} clique sets, Equation (3.20) becomes:

$$P_{G,\Gamma}(w|q) \propto \gamma_{T_{\text{syn}}} s_{\text{syn}}(q, w) + \gamma_{T_{\text{par}}} s_{\text{par}}(q, w), \tag{3.21}$$

Set	Description
T_{par}	Set of cliques containing the vocabulary node and exactly one query term node and the expansion term (w) node.
T_{syn}	Set of cliques containing the vocabulary node and exactly one query term node and the expansion term (w) node, with query term node and expansion term node connected by an edge.

Table 3.1: Summary of TE clique sets to be used.

where $\gamma_{T_{\text{syn}}}, \gamma_{T_{\text{par}}} \in [0, 1]$ and $\gamma_{T_{\text{syn}}} + \gamma_{T_{\text{par}}} = 1$. By normalising the distribution and replacing $\gamma_{T_{\text{syn}}}$ and $\gamma_{T_{\text{par}}}$ with a single interpolation parameter, γ, the rank equivalent estimate in Equation (3.21) can be rewritten as:

$$P_{G,\Gamma}(w|q) = \frac{1}{Z_\Gamma} \left[(1-\gamma) s_{\text{syn}}(q, w) + \gamma s_{\text{par}}(q, w) \right], \tag{3.22}$$

where $\gamma \in [0, 1]$, mixes the amount of syntagmatic and paradigmatic features used in the estimation, and $Z_\Gamma = \sum_{w \in V} \left[(1-\gamma) s_{\text{syn}}(q, w) + \gamma s_{\text{par}}(q, w) \right]$, is used to normalise the distribution.

higher-order TE variants

For some linguistic tasks, the similarity between a n-tuple ($n > 1$), and another word or another n-tuple ($n > 1$) may be required. For example, in comparing the similarity in meaning of two phrases, such as those used in analogy making. The formal model in Equation (3.22) can be generalised to support modelling meaning between n-tuples and words, or n-tuples with n-tuples. To illustrate the former case, the undirected graph in Figure 3.3 can be modified to replace q with a multi-word concept (n-tuple) c, as depicted in Figure 3.4.

Using the graph in Figure 3.4, the conditional probability of observing term w, given some multi-term concept c, becomes:

$$P_{G,\Gamma}(w|c) = \frac{1}{Z_\Gamma} \left[(1-\gamma) s_{\text{syn}}(c, w) + \gamma s_{\text{par}}(c, w) \right], \tag{3.23}$$

where $\gamma \in [0, 1]$, mixes the amount of syntagmatic and paradigmatic features used in the estimation, and $Z_\Gamma = \sum_{w \in V} \left[(1-\gamma) s_{\text{syn}}(c, w) + \gamma s_{\text{par}}(c, w) \right]$, is used to normalise the distribution. The information required to model the syntagmatic and paradigmatic associations between c and w can be found within the n-tuple co-occurrence information stored in an n^{th}-order TE model.

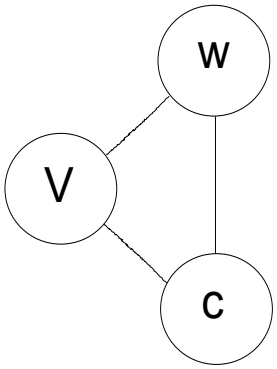

Figure 3.4: Example Markov random field for a higher-order TE model.

Equation (3.23) allows the combination of measures modelling the syntagmatic and paradigmatic associations of n-tuples ($n > 1$) and single words to be formalised, and is in line with structuralist theories that generalise the differential view of meaning to linguistic concepts, including words or phrases.

Formalism for Abstract Concepts

Section 3.2.1 just outlined how the TE framework can be extended with little change to the formalism, to cater for determining similarity in meaning between a single word and an n-tuple ($n > 1$). This result can be further generalised, not just to estimating the similarity in meaning between two n-tuples ($n > 1$), but also any *concepts* within a domain that displays syntagmatic and paradigmatic associations. For example, the interactions between social media users on the Internet has been shown to exhibit both first order (syntagmatic) and second order (paradigmatic) associations between users, from which meaningful information can be extracted for use in e-commerce activities [Yang et al., 2012]. More discussion surrounding this point will be provided in Section 8.4.2.

Until this point in the dissertation, the term *concept* has referred to a *linguistic* concept. However, the following generalised form of the TE model, relates not only to linguistic concepts but also abstract concepts.

An undirected graph that may depict relationships between two abstract concepts, c_1 and c_2, that belong to some universe of concepts U is depicted in Figure 3.5.

Using the graph in Figure 3.5, a general form of the TE model of meaning can be developed

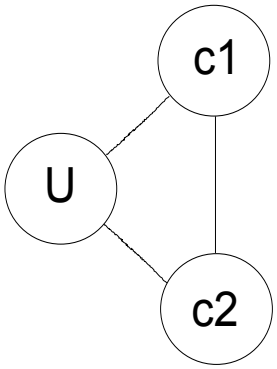

Figure 3.5: Example Markov random field for the general form of the TE model.

that estimates the conditional probability of observing concept c_2 given concept c_1:

$$P_{G,\Gamma}(c_2|c_1) = \frac{1}{Z_\Gamma}\left[(1-\gamma)s_{\text{syn}}(c_1,c_2) + \gamma s_{\text{par}}(c_1,c_2)\right], \tag{3.24}$$

where c_1 and c_2 represent concepts within a given domain that supports the existence of syntagmatic and paradigmatic associations between concepts to varying degrees. Within the TE model, Equation (3.24) effectively measures the similarity in meaning between concept c_1 and c_2.

3.2.2 Modelling Syntagmatic and Paradigmatic Associations

Now that a formal framework for combining explicit measures of syntagmatic and paradigmatic associations has been constructed a discussion of the possible measures that can be put into the framework will be provided. It is worth noting that the general framework provided in Equation (3.24) does not rely on using an SSM to provide these measures of $s_{\text{syn}}(.)$ and $s_{\text{par}}(.)$. In fact they may be provided using some other source of semantic information, including an external linguistic resource. However, for this research, we will focus on how best to find *sufficient* measures of each association using information in the representations created from the unique binding operation and TMC process developed earlier in this chapter. This decision was made so that the specific hypotheses proposed in this dissertation could be rigorously tested.

As the unique binding process presented in this work results in representations that contain explicit co-occurrence frequencies, information theoretic measures, like *mutual information* and *Kullback-Leibler (KL) divergence* can be used in addition to geometric measures. Access to both mathematical frameworks is important given recent research has shown their combination can

provide superior performance and robustness on semantic tasks when compared to traditional SSMs [Bullinaria and Levy, 2007]. This access along with the efficiency provided by the binding and compression processes developed earlier in this chapter adds weight to the decision to use these tensor representations, as opposed to representations produced by other corpus-based SSMs.

To determine which measures may effectively estimate the strength of syntagmatic or paradigmatic associations a discussion relating to the types of associations modelled by a number of popular measures is provided. Many measures of distributional and semantic similarity have been presented and studied in the past [Lin, 1998, Weeds, 2003], however, our analysis is focused primarily on understanding how syntagmatic and paradigmatic relationships may be modelled within the estimation process of several popular measures.

Cosine Measure of Semantic Similarity

The cosine metric has been shown to be a robust measure of semantic similarity used within corpus-based SSM research [Bullinaria and Levy, 2007, Landauer and Dumais, 1997, Rohde et al., 2006]. When dealing with vectors, this metric effectively computes the normalised similarity between vector $\vec{a} = (a_1, \ldots, a_n)$ and $\vec{b} = (b_1, \ldots, b_n)$, and is often expressed as:

$$\cos\theta = \frac{\langle \vec{a}, \vec{b} \rangle}{||\vec{a}||.||\vec{b}||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}, \tag{3.25}$$

where $\langle \vec{a}, \vec{b} \rangle$ is the scalar product of the vectors, $||\vec{a}||$ is the magnitude of $\vec{a}$, $||\vec{b}||$ is the magnitude of $\vec{b}$, n is the dimensionality of the vectors, θ is the angle between the vectors, and $\cos\theta \in [0, 1]$.

For vectors containing word co-occurrence information, as in HAL-based models, the cosine measure is greater for words that occur near the same words within natural language text, as indicated by the expression in the numerator of Equation (3.25). The strength of paradigmatic associations are proportional to the tendency of two words to be seen near the same words. So the cosine measure could be argued to measure the strength of paradigmatic associations in this case. SSM researchers have found that effective paradigmatic associations can be modelled with the cosine measure when a narrow context window is used to build the context vectors [Bullinaria and Levy, 2007, Sahlgren et al., 2008]. However, as the length of the sliding context window widens, the effective modelling of paradigmatic associations appears to reduce and information about syntagmatic associations is often increased.

This is also true when the cosine measure is used to determine semantic similarity between vectors in the BEAGLE and RI models, as they too are built from co-occurrence patterns [Sahlgren, 2006]. This suggests that in a co-occurrence context, the mix of syntagmatic and paradigmatic information computed by the cosine measure appears to vary depending on the length of the sliding context window used to build the context vectors, with short and long context windows best suited to modelling paradigmatic and syntagmatic associations respectively.

It is worth noting that in the case of LSA, when a document context is used, the context window is effectively the length of the document and the full word-document matrix would appear to model more information about syntagmatic associations between words. However, it has been shown that the reduced matrix created by the SVD process groups terms with strong paradigmatic associations together in the space [Sahlgren, 2006]. Therefore, the cosine of latent concepts in the reduced space effectively models information about paradigmatic associations.

This discussion suggests that within the TE model's semantic space, the cosine measure can be used to model either association, with the emphasis changing based on the length of the context window used in the vocabulary building process. This leads to the hypothesis that if relying solely on the cosine measure, effective modelling of syntagmatic and paradigmatic associations would require two separate semantic spaces, one built with a wide context window and the second with a narrow context window, respectively. This would be practical if small enough storage vectors could be used and provide satisfactory task performance, given the low computational complexity of the TE binding process and TMC technique. However, to further determine whether the cosine measure is the best choice for modelling both syntagmatic and paradigmatic associations within the TE framework, the time complexity of the measure needs to be considered.

Time Complexity: From Equation (3.25) the time complexity of the cosine measure is linear with n, the size of the vectors. In HAL-based models this can be quite large $\approx 100,000$ [Bullinaria and Levy, 2007]. Within BEAGLE and RI this is normally around 2,000-5,000 [Sahlgren et al., 2008], and within LSA this can be relatively small (i.e., around 300 [Landauer and Dumais, 1997]).

To illustrate how the TE model's tensor representations support an efficient implementation of the cosine measure consider an example using the second-order TE model. The example

uses a generalised case in which similarity between a sequence of words is measured with a word in the vocabulary. This is done to allow the cosine measure developed here to be used on applications where a sequence of words are used to prime vocabulary terms, such as in the case of query expansion within the information retrieval process, which will be investigated in Part II (Applications) of this thesis. The result can then be easily simplified to the case of measuring similarity between two words in the vocabulary by letting the sequence be of length one.

For the second-order TE model, the cosine measure for two memory matrices, $\boldsymbol{M_Q}$ and $\boldsymbol{M_w}$, where $Q = (q_1, \ldots, q_p)$ is an ordered set of terms, $M_Q = \sum_{i=1}^{p} M_{q_i}$, and $q_1, \ldots, q_p$ and w are words in the set of vocabulary terms created from the training documents, can be computed as:

$$\cos(M_Q, M_w) = \frac{\sum_{j\in\{V|w\in Q\}} f_{jw}^2 + \sum_{j\in\{V|j\neq w, w\in Q\}} f_{wj}^2 + \sum_{i\in\{Q|i\neq w\}} (f_{iw}^2 + f_{wi}^2)}{\sqrt{\sum_{i\in Q}\left[\sum_{j\in V} f_{ji}^2 + \sum_{j\in\{V|j\neq i\}} f_{ij}^2\right]}\sqrt{\sum_{j\in V} f_{jw}^2 + \sum_{j\in\{V|j\neq w\}} f_{wj}^2}}, \tag{3.26}$$

where f_{ab} is the co-occurrence frequency of word a appearing before word b in the vocabulary, and V is the set of vocabulary terms. The proof of this result is provided in Appendix A.

The denominator in Equation (3.26) is a normalising factor and helps reduce the influence of terms that occur often in the training corpus. However, the first two expressions in the numerator promote terms in Q, while the latter two promote terms that co-occur often with terms in Q. Therefore, it would seem this measure is more likely to be useful for primarily modelling syntagmatic associations between $q_1, \ldots, q_p$ and w.

The time complexity of this measure would appear to be $T(n) = O(|V||Q|)$, where $|V|$ the size of the vocabulary and $|Q|$ is the length of the sequence of terms. However, the storage vectors within the TE model hold a maximum of $\frac{D_{SV}}{2}$ (T CF) pairs, where D_{SV} is the dimensionality of the storage vectors (refer to Section 3.1.4). This means that the measure in Equation (3.26) does not need to be computed for all vocabulary terms, only those that occur with the terms in Q, and hence has a maximum time complexity when the storage vectors for the terms in Q are full, giving $T(n) = O(\frac{D_{SV}}{2}.|Q|)$.

An additional saving when computing the cosine scores is gained by noting that the numerator in Equation (3.26) will only be non-zero if word w has at least one interaction with a term in Q, or is a member of Q itself (i.e., $w \in Q$). Therefore, Equation (3.26) will only need to be computed for terms that co-occur with any of the terms in Q, i.e., those terms that have a (T CF) pair within the storage vectors of the terms in Q. All other terms in the vocabulary would have

a cosine value of zero, and hence have effectively no syntagmatic associations with the terms in Q.

The Syntagmatic Measure for a Single Target Term: To illustrate how the general case, shown in Equation (3.26), can be expressed when Q is a single word (i.e., $Q = (q)$), the measure simplifies to:

$$cos(M_q, M_w) = s_{\text{syn}}(q, w) = \frac{\sum_{j \in \{V | j \neq w, w=q\}} (f_{jq}^2 + f_{qj}^2) + (f_{qw}^2 + f_{wq}^2)}{\sqrt{\sum_{j \in V} f_{jq}^2 + \sum_{j \in \{V | j \neq q\}} f_{qj}^2} \sqrt{\sum_{j \in V} f_{jw}^2 + \sum_{j \in \{V | j \neq w\}} f_{wj}^2}}, \tag{3.27}$$

and has linear time complexity stated as $T(n) = O(\frac{D_{SV}}{2})$.

Pointwise Mutual Information

Pointwise mutual information (PMI) is an information theoretic measure [Shannon et al., 1948] that indicates the likelihood that one word will appear near another by comparing the actual conditional probability $p(w|q)$ for word q with the average or expected probability $p(w)$, and is often expressed as:

$$i(w, q) = \log \frac{p(w|q)}{p(w)} = \log \frac{p(w, q)}{p(q)p(w)}, \tag{3.28}$$

where $p(w, q) = \frac{n(q,w)}{NW}$, $n(q, w)$ is the number of times words w and q co-occur within the context window, $n(w)$ is the corpus frequency of word w, N is the total number of words in the corpus and W is the length of the sliding context window (window radius). Negative PMI values indicate that q and w co-occur with each other less than expected, and positive values indicates higher than average co-occurrence. When used with a sufficiently large context window, PMI effectively models syntagmatic associations.

Time Complexity: When implemented within the TE model's semantic space, the worst case time complexity of the PMI measure is linear with the dimensionality of the storage vector, i.e., $T(n) = O(\frac{D_{\text{SV}}}{2})$, as it requires the $p(w, q)$ value to be found in the storage vector of word q (or word w).

Cosine of Positive Pointwise Mutual Information scores

Research evaluating the effectiveness of various measures of semantic similarity has shown that the use of probabilistic based measures in combination with geometric measures can provide superior task performance. Bullinaria and Levy [2007] found that when the *positive pointwise mutual information* (PPMI) score is used to prime vector representations of words superior effectiveness across a wide variety of tasks can be achieved. This PPMI measure was defined as:

$$r(w,q) = \begin{cases} \frac{p(w,q)}{p(q)p(w)} & \text{where } \frac{p(w,q)}{p(q)p(w)} > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{3.29}$$

where $p(w,q) = \frac{n(w,q)}{NW}$, $p(q) = \frac{n(q)}{NW}$, and $p(w) = \frac{n(w)}{NW}$.

Bullinaria and Levy [2007] demonstrated that for a HAL-based model, combining the cosine measure (i.e., a geometric measure of paradigmatic associations, when a small context window is used) and PPMI (i.e., probabilistic measure of syntagmatic associations) superior performance can be achieved on a number of semantic tasks. This result enhances the intuition, previously expressed by Turney and Pantel [2010], that a framework which uses a combination of geometric and probabilistic measures appears to help improve performance of SSMs.

More importantly for this work, it also supports the premise that combining syntagmatic and paradigmatic information can improve the ability to make semantic judgements about words. A possible drawback of this PPMI approach is that the mix of syntagmatic and paradigmatic information is not formally controlled. The only way to vary it appears to be by changing the length of the underlying context window. This means the mix is determined at the point of building the space and hence to mix these information types differently the space would need to be rebuilt, which costs significant time or storage space if all possibilities are built before hand.

This cosine based PPMI model would appear to be a good benchmark model, outside of BEAGLE and PRI, as it effectively uses both syntagmatic and paradigmatic associations and has been shown to be a robust measure across a range of semantic tasks. A comparison with this approach may help identify the benefits gained by using the TE model's formal framework to explicitly mix information about syntagmatic and paradigmatic associations.

Time Complexity: Since the cosine of the PPMI measure combines both syntagmatic and paradigmatic information to some degree it is not a good choice to use as either the syntagmatic or paradigmatic measure within the TE model. However, PPMI on its own may be an effective variant of PMI that models the strength of syntagmatic associations well, and also has linear time complexity of $T(n) = O(\frac{D_{\text{SV}}}{2})$.

A Novel Paradigmatic Measure

Many measures of semantic similarity based on either distributional similarity [Lin, 1998] or geometric distances between representations [Schütze and Pedersen, 1993] capture some mix of syntagmatic and paradigmatic information. This mix appears to be sensitive to the context window used to build the representations. This is due to the fact that within natural language syntagmatic associations exist between words found quite far apart [Xu and Croft, 1996], while paradigmatic associations are best modelled by looking at close neighbouring words [Bullinaria and Levy, 2007].

However, given the cosine measure developed in Equation (3.26) can be argued to better model syntagmatic association when a large context window is used within the second-order TE model, but does not appear to be well suited to modelling paradigmatic associations even when a narrow context window is used, there exists an opportunity to develop a new measure specifically adapted to modelling the strength of paradigmatic associations between tensor representation within the TE model.

To create a measure of paradigmatic association between two words q and w, the score for words that co-occur with the same words as q should be enhanced, irrespective of whether q and w co-occur themselves. A probabilistic measure that achieves this, can be expressed as:

$$s_{\text{par}}(q, w) = P(w|q) = \frac{1}{Z_{\text{par}}} \sum_{i \in V} \frac{f_{iq} f_{iw} + f_{qi} f_{wi}}{f_q f_w}, \tag{3.30}$$

where f_q is the vocabulary frequency of term q, f_{qi} is the ordered co-occurrence frequency of term q before term i, V is the set of terms in the vocabulary, and $Z_{\text{par}} = \sum_{w \in V} \left[\sum_{i \in V} \frac{f_{iq} f_{iw} + f_{qi} f_{wi}}{f_q f_w} \right]$ normalises the distribution. A more efficient rank preserving form of Equation (3.30) can be stated as:

$$s_{\text{par}}(q, w) \propto \sum_{i \in V} \frac{f_{iq} f_{iw} + f_{qi} f_{wi}}{f_q f_w}. \tag{3.31}$$

This measure provides a score indicating the relative strength of paradigmatic associations

between w and q, irrespective of the strength of syntagmatic associations between w and q and the length of the context window. It is worth noting that I am not claiming that **no** syntagmatic information will be displayed by the results returned by this measure, but more just that in the absence of syntagmatic associations between q and w, if they share any common neighbours (within the context window) they will get a paradigmatic score greater than zero.

Time Complexity: Initially, the time complexity of Equation (3.31) appears polynomial, with $T(n) = O(|V|^2)$, where $|V|$ is the number of terms in the vocabulary. However, at an implementation level, the co-occurrence information for terms q and w are accessed from their respective storage vectors, which are of a fixed dimension, D_{SV}. Since each storage vector in the second-order TE model holds a maximum of $\frac{D_{SV}}{2}$, (T CF) pairs, the worst case time complexity of this paradigmatic measure is actually $T(n) = O(\frac{D_{SV}^2}{4})$. Figure 3.2 showed that for the task of synonym judgement and a $CFC = 0.002$, a $D_{SV} = 250$ can provide optimal performance, hence $T(n) = 15,625$. This result means that within the TE model, the paradigmatic measure can provide optimal task effectiveness on synonym judgement for the same time complexity that a linear order measure working with context vectors of 15,625 dimensions could produce. This is better than HAL-based models that use around 100,000 dimensions, however, when compared to RI this is not so flattering.

This result highlights the time complexity of the paradigmatic measure as its weakness, and the size of the storage vectors used to underpin it will need to be considered if this measure is chosen as the preferred, due to its robustness at modelling paradigmatic associations for various context window sizes.

Summary

It is acknowledged that past research appears to suggests that for most effective modelling of paradigmatic and syntagmatic associations a narrow and wide context window should be used, respectively. This means that to best model syntagmatic and paradigmatic associations within the TE model two different semantic spaces may be required, especially if a single similarity measure is used. If this were the case, then this would impact the computational complexity of the TE model, with the memory footprint and time required to build the space doubling in order. Given the computational complexity of the process used to build the semantic space within the TE model, outlined in Section 3.1.4, even doubling these complexities puts the TE model on

par with the permuted random indexing (PRI) model, which is more efficient than most SSMs, including BEAGLE, HAL-based models and LSA. However, given the paradigmatic measure developed in Equation (3.30) primarily models paradigmatic information, irrespective of the length of the context window, it can be argued that basing both the syntagmatic and paradigmatic measure of the one space using an average length context window may still provide effective modelling of both linguistic associations.

Both cases: (i) using one space to underpin both measures, and (ii) using two separate spaces to underpin the measures within the TE model will be investigated within this work.

The analysis of similarity measures and their ability to model syntagmatic and/or paradigmatic associations has focused on a few popular measures, as well as presenting a novel measure of paradigmatic information. Further analysis could be undertaken, however, for practical reasons, and given the primary goal is to demonstrate the performance when both syntagmatic and paradigmatic information are explicitly modelled and combined in a controlled way, the following measures were chosen for the initial evaluation of the TE model, to be carried out in Chapter 4:

1. **Syntagmatic Associations**: The efficient *syntagmatic* measure developed in Appendix A, and expressed in Equation (3.27) was chosen as the primary measure for modelling syntagmatic associations within the TE framework. It was chosen over PPMI as it uses word order information within the calculation. It is hypothesised that the inclusion of word order information gives the model more flexibility and allows it to be applied on a greater range of tasks, including word priming tasks, as demonstrated in Section 3.2.3.

2. **Paradigmatic Associations**: The novel paradigmatic measure developed in this research, and expressed in Equation (3.30), will be used as the primary measure for modelling paradigmatic information. This was chosen over the cosine measure using a smaller context window due to its ability to model paradigmatic information irrespective of the context window length used to build the semantic space. However, the impact of the size of the storage vectors on its computational complexity will need to be managed given the discussion in Section 3.2.2.

Combining Equation (3.27) and Equation (3.30) into the formal TE framework expressed in Equation (3.22), produces the following rank equivalent estimate for the similarity in meaning

of word q and w, with respect to their meanings as defined by structural linguistics:

$$P_{G,\Gamma}(w|q) \propto (1-\gamma)\frac{\sum_{j\in\{V|w=q\}} f_{jq}^2 + \sum_{j\in\{V|w=q,j\neq w\}} f_{qj}^2 + (f_{qw}^2 + f_{wq}^2)}{\sqrt{\sum_{j\in V} f_{jq}^2 + \sum_{j\in\{V|j\neq q\}} f_{qj}^2}\sqrt{\sum_{j\in V} f_{jw}^2 + \sum_{j\in\{V|j\neq w\}} f_{wj}^2}} + \gamma\sum_{i\in V}\frac{f_{iq}f_{iw} + f_{qi}f_{wi}}{f_q f_w}, \tag{3.32}$$

where $\gamma \in [0,1]$, mixes the amount of syntagmatic and paradigmatic information used in the estimation. Using the time complexity analysis for each measure provided earlier, the worst case time complexity of the TE model in Equation (3.32) is based on the paradigmatic measure, and hence is $T(n) = O(\frac{D_{SV}^2}{4})$.

The general form of the TE model, shown in Equation (3.24), allows the similarity between abstract concepts to be estimated based on their syntagmatic and paradigmatic associations. The measures chosen to model syntagmatic and paradigmatic associations can also be used in the general form of the TE model. The resulting expression estimates the similarity in meaning of concepts c_1 and c_2 with respect to their meanings as defined within structural linguistic theory, and can be expanded to:

$$P_{G,\Gamma}(c_2|c_1) \propto (1-\gamma)\frac{\sum_{j\in\{U|j\neq c_2,c_2=c_1\}}(f_{jc_1}^2 + f_{c_1j}^2) + (f_{c_1c_2}^2 + f_{c_2c_1}^2)}{\sqrt{\sum_{j\in U} f_{jc_1}^2 + \sum_{j\in\{U|c_j\neq c_1\}} f_{c_1j}^2}\sqrt{\sum_{j\in U} f_{jc_2}^2 + \sum_{j\in\{U|j\neq c_2\}} f_{c_2j}^2}} + \gamma\sum_{i\in U}\frac{f_{ic_1}f_{ic_2} + f_{c_1i}f_{c_2i}}{f_{c_1}f_{c_2}}. \tag{3.33}$$

It is proposed that Equation (3.33) provides an expression that can be used for comparing the meaning of abstract concepts within any domain that exhibits syntagmatic and paradigmatic associations. Example domains where this may be the case are presented in Section 8.4.2. It is important to note that the measures chosen aim to model predominantly syntagmatic or paradigmatic associations. However, this does not mean they do not also model in some form other associations. Therefore, a practical choice is made to initially implement these measures within the TE model's framework. Changes and enhancements to these measures are developed and evaluated later in this work.

3.2.3 Evaluating the Measures of Syntagmatic and Paradigmatic Association

To gain an intuitive understanding of how well each measure models the word associations they are claimed to, it is worthwhile evaluating their effectiveness on a number of small word

association tasks. Many approaches for evaluating the effectiveness of lexical distributional similarity measures have been proposed by *natural language processing* (NLP) researchers in the past [Weeds, 2003], including the use of gold standard tests, comparison to human plausibility judgements, application-based evaluation tasks and through the use of manually constructed lexical resources.

These techniques are often aimed at evaluating the measures ability to achieve a given lexical result, i.e., to list the hypernyms for a given word based on the synsets in WordNet (Section 2.5.3) or produce words with a similar part of speech as the target word; and rely on knowing the part of speech of the words involved. Even though these lexical results may rely on syntagmatic and paradigmatic information in some form, they are often very restrictive in their definitions, such as an effective paradigmatic measure should produce the hypernyms in a WordNet synset, irrespective of the document collection used to base the measure. This deterministic approach does not reflect the probabilistic nature of word associations modelled within corpus-based representations.

Further, the measures used in this research are designed to provide an indication of the strength of syntagmatic and paradigmatic associations between words *based on the distributional evidence*. This means that the TE model of word meaning aims to induce meaning based solely on distributional evidence, irrespective of external linguistic resources, including information about part of speech. Therefore, a more probabilistic model of associations will result, and hence a more intuitive approach to evaluating the measures is justified. To achieve this a word association task, similar to that undertaken in Rapp [2002] and a word priming task, as performed in Jones and Mewhort [2007] are used to initially evaluate the intuitive effectiveness of the measures.

These two experiments will use the *Touchstone Applied Science Associates*, Inc. (TASA) corpus to build up the representation within the TE model. The TASA corpus is a 44,486 document collection of text representative of the reading material that a person is supposed to have been exposed to by their first year in college.

Modelling Syntagmatic and Paradigmatic Associations

To get an intuitive feel for the types of associations being modelled by the selected measures, and the impact on them when computed from the same semantic space as opposed to two

	Syntagmatic: $s_{\text{syn}}(q, w)$		Paradigmatic: $s_{\text{par}}(q, w)$		
	$cwr = 5$ $D_{sv} = 500$	$cwr = 10$ $D_{sv} = 1,000$	$cwr = 5$ $D_{sv} = 500$	$cwr = 2$ $D_{sv} = 100$	$cwr = 5$ $D_{sv} = 100$
Test word: q					
heart	blood attack disease attacks muscle	blood attack disease muscle soul	sleep age love trouble arm	nose legs parents friends hands	family wife hand arm brother
cold	**hot** weather winter **warm** air	**hot** **warm** winter weather air	inside girls leave feel dead	**hot** rain angry afraid **warm**	key play mouse hold feel
swim	fish water sharks ocean underwater	fish water ocean underwater sharks	jump catch lake kids boat	hit fly stand watch jump	jump afraid carry alone explain
apple	tree pie trees orchard blossoms	tree pie orchard trees blossoms	log feathers basket grandpa cookies	cookies listening sandwiches digging painting	cookies lunchbox sandwiches hearing christmas

Table 3.2: Comparison of syntagmatic and paradigmatic associations for 4 test words using the syntagmatic measure in Equation (3.27) and paradigmatic measure in Equation (3.31), respectively. Items in **bold** indicate words that overlap between syntagmatic and paradigmatic associations. cwr and D_{sv} indicate the context window radius and dimensionality of the storage vectors, respectively.

separate semantic spaces (i.e., built using two different length context windows), the 5 words with the strongest syntagmatic and paradigmatic associations for 4 test words are listed in Table 3.2. This sort of intuitive investigation has been used in the past to gauge how well each measure models syntagmatic and paradigmatic associations, and considers the extent of overlap between the measures [Rapp, 2002].

Table 3.2 reports the five words with the strongest syntagmatic associations (in column 2 and 3) and strongest paradigmatic associations (in column 4,5 and 6) for each test word. These results demonstrate that the syntagmatic and paradigmatic measures model quite different types of associations. In fact, the only test word were overlap between association types exists is for the word *cold*, where the words *hot* and *warm* are listed in both syntagmatic and paradigmatic lists. Rapp [2002] also found this overlap for the test word *cold*.

This overlap may exist because of the frequent use of either *hot* or *warm* in the same sentence as *cold*, such as: *the hot and cold taps*; as well as in contexts conducive to building strong paradigmatic associations, such as: *turn the cold water on* or *turn the hot water on*. These cases may be more common among adjectives, as seen by the lack of overlap between syntagmatic and paradigmatic association lists for the test words that are not adjectives.

Another interesting observation is that the within-association overlap for the syntagmatic lists is quite high. Indicating there is little difference in modelling syntagmatic associations when the context window radius is changed from 5 to 10, and the storage vector dimensionality changed from 500 to 1000. However, within the lists of paradigmatically associated words there appears to be much more variation between the three unique combinations of context window radius and storage vector dimensionality parameter values. This may indicate that paradigmatic associations are more sensitive to these model parameters.

To gauge the intuitive effectiveness of the syntagmatic measure, we could simply ask: *Is it highly likely that the suggested words in columns 2 an 3 would appear near the test word in a natural language sentence?* Using this question on the chosen test words, the syntagmatic measure would appear to be intuitively effective.

However, testing the intuitive effectiveness of the paradigmatic measure is not as straight forward. A suggested test question may be: *Within an example sentence, are the words listed in the paradigmatic columns likely to be seen in place of the test word or near the words suggested in the syntagmatic columns?*

To demonstrate, consider the test word *swim*: *jump* has a strong paradigmatic associations with *swim*, as indicated by its presence in columns 4, 5 and 6. Using the proposed test process, and the following example sentences: (i) *the fish* ***swim/jump*** *in/(out of) the water/ocean*, and (ii) *kids* ***swim/jump*** *in the water*; it would seem reasonable to suggest that *jump* may have a strong paradigmatic association with the test word *swim*.

Based on these findings, it can be argued that the chosen measures of syntagmatic and paradigmatic associations are intuitively effective.

Word Priming using Syntagmatic Information

To demonstrate the importance and effectiveness of word order information that is encoded into the TE model's semantic space, an evaluation of the syntagmatic measure's ability to estimate the most likely word to precede or succeed a target word can be performed. To achieve this in a second-order TE model, co-occurrence frequencies in the direction of interest can be isolated, by ignoring elements on the row or column *not* of interest in the memory matrices. For example, when considering the word most likely to precede a target term all element values on row q of M_q (except for column q) can be ignored, and for the most likely succeeding word, all element values on column q would be ignored. The same is done for the memory matrix of word w.

At an implementation level, this would be done by examining the sign of the term-id (T) component of the (T CF) pair in the storage vectors. A negative sign in front of the term-id would indicate a co-occurrence before the target term, and a positive sign would indicate a succeeding co-occurrence. For estimating the words most likely to precede a target term, use only the pairs with a negative term-id in the calculations. For estimating the terms most likely to succeed a target term, use only pairs with positive term-id values.

To illustrate algebraically how the syntagmatic measure in Equation (3.27) can be used to compute the most likely preceding word w for word q, the terms relating to succeeding co-occurrences for both words are removed from the equation. The resulting measure becomes:

$$s_{\text{syn}_{\text{pr}}}(q, w) = \frac{\sum_{j \in \{V|q=w\}} f_{jw}^2 + f_{wq}^2}{\sqrt{\sum_{j \in V} f_{jq}^2} \sqrt{\sum_{j \in V} f_{jw}^2}}, \tag{3.34}$$

with an equivalent expression, using f_{qw} instead of f_{wq} in the numerator and qj and wj in the denominator, created to calculate the most likely succeeding terms. Table 3.3 provides a list of most likely preceding and succeeding terms produced by the second-order TE model for a list

of target words identified in [Jones and Mewhort, 2007] for the BEAGLE model. The results illustrate the influence of the asymmetric nature of the memory matrices, and the effectiveness of the cosine measure to identify the strongest ordered syntagmatic associations.

KING		PRESIDENT		WAR	
____ king	king ____	____ president	president ____	____ war	war ____
luther	jr	vice	roosevelt	civil	ii
martin	midas	elected	kennedy	world	ended
dr	arthur	former	nixon	revolutionary	effort
french	minos	new	johnson	spanish-american	began
rex	queen	our	lincoln	during	between

Table 3.3: Top 5 preceding and suceeding words produced by the syntagmatic measure.

3.2.4 Summary

These two tasks demonstrate that the representations created by the TE model's binding process, and the choice of syntagmatic and paradigmatic measures made in Section 3.2.2 appear to intuitively model the two associations argued by structural linguistics to be at the basis of word meanings. However, a comparison with benchmark models on a wide range of tasks is needed to rigorously evaluate the effectiveness of the TE model.

Chapter 4

Evaluating the Tensor Encoding (TE) Model

4.1 Overview

The story so far has focused on the evolution of SSMs, particularly the use of fixed dimension approaches to enhance efficiency, the encoding of word order to reduce the incidence of non human-like errors, and the use of tensor representations to increase the range of tasks to which SSMs can be applied. The TE model developed in Chapter 3, levered these evolutionary advances in SSM technology to build tensor representations to form the foundations of a formal model of word meaning grounded in structural linguistics.

A computational analysis of the TE model's vocabulary building process and measures of semantic similarity suggested that the model would be more efficient than most corpus-based SSMs (Section 3.1.4 and Section 3.2.2). Initial experiments using two measures, one of syntagmatic and the other paradigmatic associations demonstrated their intuitive effectiveness (Section 3.2.3). However, to evaluate whether the TE model, using these measures, can achieve superior effectiveness when compared to current corpus-based SSMs, an evaluation of its performance on a number of benchmark tasks is required.

Recent work evaluating corpus-based SSMs has seen robust investigation into the impact of model parameters and choice of similarity measures on model effectiveness [Bullinaria and Levy, 2007]. This work identified the cosine of PPMI vectors as the strongest performing measure, and also provided a set of semantic tasks on which to evaluate SSMs. The three tasks common to these works include: (i) a synonym judgement task, (ii) a semantic distance task, and (iii) a semantic categorization task. The effectiveness of the TE model against the HAL based model using the cosine of PPMI vectors will be evaluated for these three tasks. Further,

a comparison of the TE model's performance on the task of synonym judgement will also be compared to that achieved by implementations of the BEAGLE and PRI model.

In addition, the area of medical retrieval is receiving growing interest within the information retrieval community. Within medical retrieval the task of detecting similarity between medical concepts has been shown to help improve search results by overcoming issues such as vocabulary mismatch. Vocabulary mismatch relates to the use of more than one term to describe the same concept across documents, such as the use of *high blood pressure* and *hypertension* within patient records. A search for documents relating to *high blood pressure* should also return documents relating to *hypertension*. Therefore, the TE model is also evaluated on a number of medical concept similarity judgement tasks using gold standard data sets.

4.2 TOEFL synonym judgement

The synonym judgement part of the Test of English as a Foreign Language (TOEFL) was first used by Landauer and Dumais [1997] and has been used extensively since to evaluate semantic space model performance on synonym judgement [Bullinaria and Levy, 2007, 2012, Jones and Mewhort, 2007, Sahlgren et al., 2008]. This researcher is grateful to Landauer for providing this data set.

In the synonym-finding part of TOEFL the participant is asked to choose one of four provided words as the most similar to the question word. For example, for the word *physician* which of the following is closest in meaning: *chemist*, *pharmacist*, *nurse*, *doctor*. It was reported that for a large sample of applicants to U.S. colleges, coming from non-English speaking countries, the average result on the synonym test was 51.6 items correct out of 80 (or 64.5%) [Landauer and Dumais, 1997].

4.2.1 Experimental Setup

Benchmark models

A review of past papers and results on the TOEFL synonym test[1] suggests that the corpus used, preprocessing of documents[2] and the resulting vocabulary size may impact the performance

[1] Listed at http://aclweb.org/aclwiki

[2] Including the use of external linguistic resources like POS taggers

achieved by any given model [Stone et al., 2008]. Therefore, comparisons of TOEFL performance between papers is unlikely to be reliable. A more robust comparison may be achieved by evaluating the models of interest on the same data configuration, hence an implementation of (i) BEAGLE based on the Jones and Mewhort [2007] paper (Section 2.6.1), (ii) the permuted RI (PRI) model based on the Sahlgren et al. [2008] paper[3] (Section 2.6.2) and (iii) the HAL-based approach (Section 2.5.3) using the cosine of PPMI measure outlined in Bullinaria and Levy [2007].

The BEAGLE [Jones and Mewhort, 2007], and permuted RI models [Sahlgren et al., 2008] *encode* more structural information and hence go beyond traditional corpus-based methods. The performance of the TE model is also compared to an implementation of a HAL based approach that uses the cosine of vector based positive pointwise mutual information (PPMI) scores to measure semantic similarity. This approach was shown to be the most effective on a number of semantic tasks, including synonym judgement, when compared to numerous other geometric and probabilistic measures [Bullinaria and Levy, 2007].

For all models, except PPMI, the vocabulary building process used a stoplist[4]. Stoplists are commonly used to remove high frequency, low information terms, often referred to as closed terms, from the vocabulary. The use of stoplists have been shown to improve the effectiveness of corpus based SSMs on a wide range of tasks [Rapp, 2003, Rohde et al., 2006]. However, Bullinaria and Levy [2012] report that their PPMI based model performed better without a stoplist. Therefore, no stoplist was used within the implementation of the PPMI model. When building representations within each of the models, sentence boundaries were ignored, along with terms containing numerics.

Corpora

To evaluate the influence of training corpus size on task effectiveness the corpora used to build the semantic spaces for the TOEFL task included the *Touchstone Applied Science Associates, Inc.*, Inc. (TASA) corpus and *British National Corpus* (BNC) which have been used extensively in past corpus-based SSM research [Bullinaria and Levy, 2007, 2012, Jones and Mewhort, 2007, Landauer and Dumais, 1997, Sahlgren et al., 2008].

[3]The permuted RI model was built using code from the semantic vectors codebase, http://code.google.com/p/semanticvectors/

[4]418 word Indri stoplist, used in many information retrieval experiments.

The smaller TASA corpus is a 44,486 document collection of text representative of the reading material that a person is supposed to have been exposed to by their first year of college. When built using the stoplist shown in Appendix B and ignoring terms containing numerics the TASA corpus contains 133,753 unique terms.

The larger BNC is a 100 million word corpus that is made up of 90% written text with the remaining 10% being transcribed spoken text. When built using the stoplist shown in Appendix B and ignoring terms containing numerics the BNC corpus contains 482,626 unique terms. The BNC corpus is supplied with part of speech tags. However, as the scope of this research does not include the use of external linguistic information these tags were removed from the corpus. Along with the fact that punctuation is also removed, this ensures the results are a conservative measure of the potential of each model.

Model parameters

To fairly compare performance on each of the models, the parameters used within each model need to be identified and controlled to ensure the influence of each variable on task effectiveness can be more rigorously evaluated.

For semantic space models, the configuration used to build representations is controlled by the length of the context window and the size of the underlying vectors. In addition, the TE model also uses the co-occurrence frequency cut-off (CFC) parameter within the TMC technique (Section 3.1.4). For the semantic tasks evaluated in this chapter the *CFC* parameter (Section 3.1.4) within the TE model was set to 0.002 to ensure the BEAGLE and PRI models used the same number of free parameters when building their representations. Setting $CFC = 0.002$ means that the oldest (T CF) pair in the storage vector of the target term (w) that has a $\frac{CF}{F_w} \leq \frac{1}{500}$ will be marked for replacing if the storage vector is full. In effect, indicating that the oldest co-occurring term that has co-occurred at most once for every 500 occurrences of the target term will be marked as having low enough informational value to be discarded from the representation.

When performing similarity calculations the TE model is able to explicitly set the mix of syntagmatic and paradigmatic information used within the estimate. This is analogous to the mix of order and context information that BEAGLE and PRI combine in their estimates. However, the cosine of the PPMI approach presented by Bullinaria and Levy [2007] does not

have a parameter to allow the mix to be modified. Therefore, BEAGLE, PRI and the TE model can be said to have an additional free parameter when compared to PPMI. The sensitivity of the TE model's effectiveness with respect to this additional mixing parameter will be evaluated for each of the tasks in this chapter.

4.2.2 Experimental Results

Sensitivity to Context Window Length

All of the models being evaluated build their lexicons by moving a context window across the underlying text. Section 3.2.2 discussed how the size of the context window impacts the ability to effectively model syntagmatic and paradigmatic associations between terms within the semantic space. Therefore, the type of information best suited to a task will influence the choice of context window length. For example, synonym judgements rely more heavily on paradigmatic associations, which are best modelled using a narrower context window.

To evaluate the impact of context window size on each model's performance, two experiments were devised.

1. Comparing fixed dimension approaches: The first experiment compared the performance of two fixed dimension models (BEAGLE and PRI) with the TE model on the TASA corpus for various context window sizes. The results are shown in Figure 4.1 and were reported by Symonds et al. [2011a]. It is worth noting that the BEAGLE and PRI scores are the average of three separate runs, as these models produce different results on each run due to the initial random assignment of environment vectors.

The BEAGLE results were similar to those reported in Jones and Mewhort [2007], with any improvement likely due to the reduced context window radius used in our experiments [5]. The performance of our PRI implementation appears to be much lower than that reported in [Sahlgren et al., 2008], possibly due to the difference in vocabulary size. Their TASA vocabulary was reduced to 74,100 terms by using stemming and high frequency cut-offs, as compared to the resulting 134,000 term vocabulary used in these experiments.

[5] The original BEAGLE research [Jones and Mewhort, 2007] used a context window size equal to the sentence length

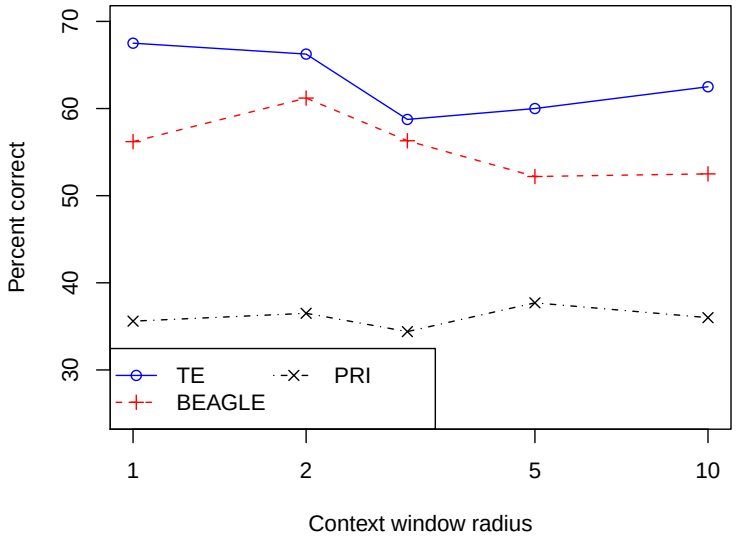

Figure 4.1: Performance of the TE, BEAGLE and PRI models on the synonym judgement part of TOEFL when trained on the *TASA* corpus for various context window radii. The evaluation was performed for the optimal context vector dimensions for all models and $\gamma = 0.1$ for the TE model.

2. Different corpus sizes: The second experiment compared the performance of the TE and PPMI models on the TASA and BNC corpora for various context window radii. BEAGLE was not included due to its excessive computational costs, and PRI was left out due to the poor effectiveness on the previous experiment. A frequency cut-off of 222 was used on the PPMI model, to ensure only the 100,000 most frequent terms were used in the similarity calculations. This approach was shown by Bullinaria and Levy [2007] to provide optimal performance in terms of vector sizes. The TE model used storage vectors with 2,000 dimensions, and a $\gamma = 0.4$.

The results (Figure 4.2) illustrate that the TE model can outperform the PPMI model on both corpora, with the best score achieved by the TE model when the context window has a radius of one. The best effectiveness of our PPMI implementation is lower than those reported in Bullinaria and Levy [2007] on the BNC corpus (i.e., 68% compared to approx 85%). This may be due to differences in their preparation of the training documents, and in TOEFL test words used, as Bullinaria and Levy [2007] report that they had made some modification to the original TOEFL test words. However, we do find that our implementation of the PPMI model achieves its best score using a context window radius of 3 on the BNC corpus, which corresponds with the findings reported in Bullinaria and Levy [2007].

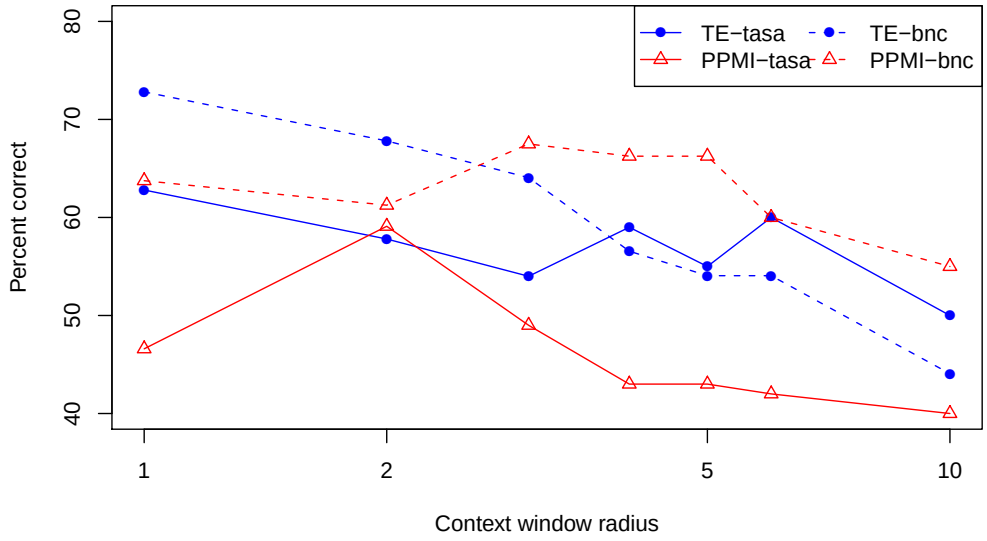

Figure 4.2: Performance of the PPMI and TE model on the synonym judgement part of TOEFL when trained on the *TASA* and *BNC* corpora for various context window radii. The evaluation was performed using a PPMI vector of 100,000 dimensions, and TE storage vector of 2,000 dimensions and $\gamma = 0.4$.

Summary

Both experiments demonstrate that the TE model can achieve superior task effectiveness over the chosen benchmark models. The second experiment demonstrates that as the size of the vocabulary increases (from TASA to BNC) the TE model's performance on TOEFL improves from around 68% to 72%, this is consistent with findings on other corpus-based approaches [Bullinaria and Levy, 2007].

In the scheme of best ever TOEFL scores achieved by a corpus-based approach, the performance of the models evaluated in this research are not outstanding[6]. However, as the goal of these experiments was to compare models on the same data set to allow the findings to be drawn into reliable conclusions, the TE model was evaluated without many of the interventions used by those top reported models, such as vocabulary downsizing through stemming or increased frequency cut-offs, and the use of external linguistic resources. Therefore, the results reported in this work can be considered to be a conservative indication of performance.

[6] http://aclweb.org/aclwiki

Sensitivity to Storage Vector Dimensionality

The other free parameter used when building the vocabulary of the TE model relates to the dimensionality of the storage vectors. The task performance for various storage vector dimensions is shown in Figure 4.3. The graph suggests that the TMC technique, used when storing the TE model's representations, appears to assist in providing superior performance with fewer dimensions, as little as 200 and 500 dimensions on the TASA and BNC training collections, respectively. This may suggest that the TMC technique is able to remove noise from the system and provide better access to information about word associations that help improve task effectiveness when performing synonym judgements.

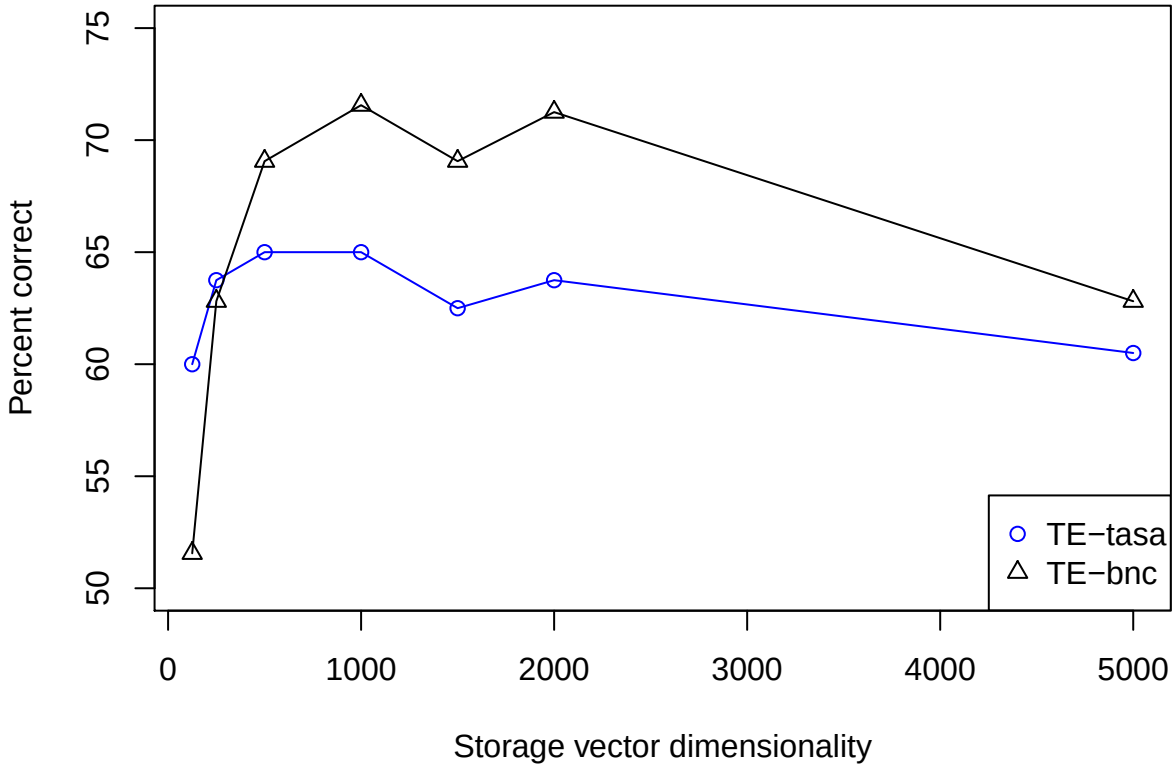

Figure 4.3: Performance of the TE model on the synonym judgement part of the TOEFL for various storage vector dimensions on both the TASA (TE-tasa) and BNC (TE-bnc) corpus. The evaluation was carried out using a context window radius of 2 and $\gamma = 0.4$.

Sensitivity to Gamma

The ability to explicitly combine syntagmatic and paradigmatic information within the TE model's formal framework, developed in Section 3.2.1, is achieved via the mixing parameter, γ in Equation (3.32). The sensitivity of the TE model's performance to this mixing parameter when performing synonym judgements is illustrated in Figure 4.4.

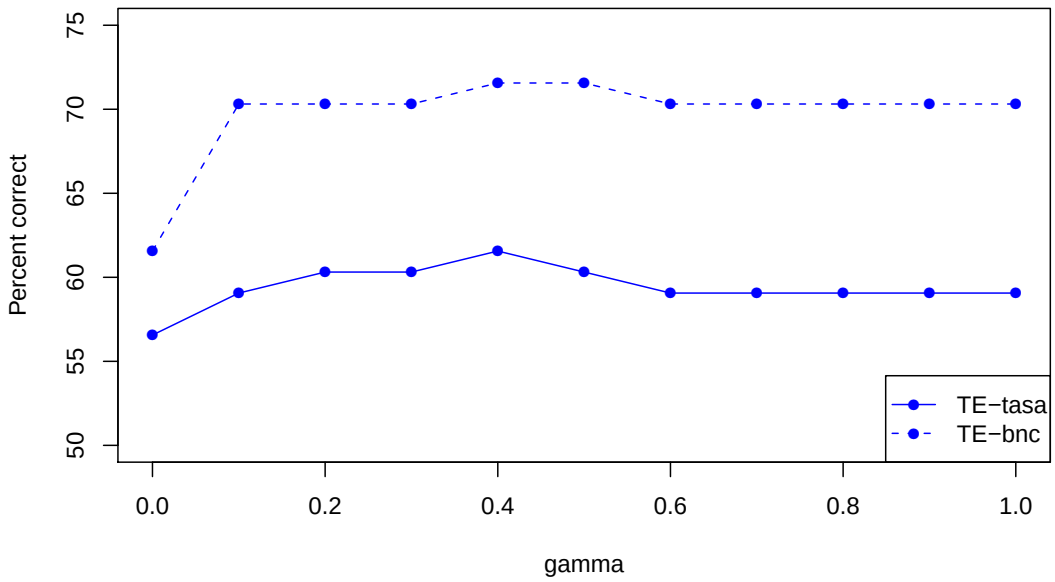

Figure 4.4: Performance of the TE model on the synonym judgement part of the TOEFL for various values of gamma (γ) on both the TASA (TE-tasa) and BNC (TE-bnc) corpus. The evaluation was carried out using a context window radius of 2 and storage vector of 1,000 dimensions.

This graph suggests that the TE model's best performance is found when some mix of syntagmatic and paradigmatic information is used. A similar result is found when using other storage vector dimensions and context window lengths. It is hypothesised that this may be due to the robustness gained by using two different sources of linguistic information to estimate the similarity of words.

Figure 4.4 also confirms the finding by other researchers that increasing the amount of training data (i.e., using the BNC rather than the TASA corpus) usually leads to consistent improvements in performance for the task of synonym judgement [Bullinaria and Levy, 2007].

Addressing weaknesses in LSA

Landauer and Dumais [1997] indicated that some of the errors produced by LSA on the TOEFL synonym judgement task, and that were not made by students, may be attributed to the fact that LSA was more sensitive to paradigmatic associations, and not syntagmatic. For example, Perfetti [1998] commented that on the TOEFL synonym judgement task, LSA chose *nurse* (0.47) over *doctor* (0.41) for the question word of *physician*.

The design of the TE model ensures sensitivity to syntagmatic and paradigmatic associations can be tuned for a given task. To compare LSA to the TE model using Perfetti's selective example, we can see that when trained on the TASA corpus and using a γ of 0.4, the TE model chose *doctor*, $P(\text{doctor}|\text{physician}) = 0.019$ over nurse, $P(\text{nurse}|\text{physician}) = 0.018$.

4.2.3 Conclusion

The above evaluation on the TOEFL synonym judgement task suggests that the TE model provides superior task effectiveness when compare to two SSMs that encode word order (BEAGLE and PRI) and a strong HAL based model.

The TE model also appears to provide this superior effectiveness at a much lower computational complexity. This is attributed TMC technique that provides information sensitive compression. The results also suggest that the encoding of word order information along with the TMC approach helps reduce non-human like errors on the task of synonym judgement.

To investigate whether these findings translate to other semantic tasks, a semantic distance and semantic categorization task will be undertaken.

4.3 Semantic Distance and Categorization Tasks

To investigate the ability to model different forms of semantic relations within the TE model, a benchmark semantic distance and semantic categorization experiment were undertaken. The results and their discussion are grouped together to provide a broader comparison of the effectiveness of the TE model compared to the PPMI benchmark model.

4.3.1 Experimental Setup

The results for the following two experiments have been reported by Symonds et al. [2012b].

Semantic Distance Task

The semantic distance task, presented in Bullinaria and Levy [2007], is a multiple choice task used to test the ability of a semantic model to detect semantic difference among more common words. Unlike the TOEFL synonym judgement test, which tests the fine distinctions between words that tend to occur infrequently within the corpus, the distance comparison task tests the large scale structure of the semantic space by comparing semantic relatedness of commonly occurring words. The task involves 200 pairs of semantically related words, such as "king" and "queen", "concept" and "thought". It is implemented by comparing the semantic relatedness, as defined by each model, of the target word to its pair and 10 other randomly chosen words in the list. The resulting performance is the percentage of control words that are further from the target than its related word.

Semantic Categorization Task

The semantic categorization task tests the ability of a semantic model to classify words within their correct semantic category. Ten words are taken from each of 70 semantic categories (e.g., fruits, sports, colours) based on human category norms [Van Overschelde et al., 2004]. The resulting performance is the percentage of the 700 test words that fell closer to their own category centre (correctly categorized) rather than another. The category centres are constructed by finding the mean of the geometric representations corresponding to the words in each category (excluding the target word under consideration).

The original semantic categorization test performed by Bullinaria and Levy [2007] used human category norms produced by Battig and Montague [1969]. An updated set of norms from Van Overschelde et al. [2004] was chosen because the original survey was over 40 years old and changes in test word meanings may have occurred over that time.

4.3.2 Experimental Results

Given the poor effectiveness of the PRI model on the synonym judgement experiment and the computational complexity of BEAGLE, the following experiments involved comparing the performance of the TE model with the PPMI model used in the previous synonym judgement experiment. The PPMI model was shown to be the strongest corpus-based model on these two tasks, when compared to the effectiveness of fourteen other models [Bullinaria and Levy, 2007].

Both the PPMI and TE models construct their semantic space by moving a triangular context window across the underlying text. A triangular context window gives terms closer to the focus term a higher weighting, with the weighting reducing with the distance from the focus term. Figure 4.5 shows how the performance of the TE and PPMI models depend on the context window radius.

The slight drop in best performance on the semantic distance task for our implementation of the PPMI model when compared to the Bullinaria and Levy [2007] implementation (92% compared to 96%) may be related to possible differences in the preparation of the training corpus and differences in the proximity scaling function used to create a triangular context window. In the experiments carried out in this work the implementations of the TE and PPMI models used the linear scaling function shown in Section 3.1.3.

The difference in best performance on the semantic categorization task for our implementation of the PPMI model when compared to the Bullinaria and Levy [2007] implementation (85% compared to 80%) is likely due in large part to our decision to use an updated set of category norms for this experiment.

However, in both experiments the sensitivity of the PPMI model's performance to the context window length corresponds with the results reported by Bullinaria and Levy [2007], which show the PPMI model achieves superior results for a much smaller context window.

Figure 4.5 also suggests that a context window radius of 4 allows the TE model to more effectively combine information about syntagmatic and paradigmatic associations. The influence of each type of information in achieving optimum effectiveness can be seen in Figure 4.6. It appears that the optimal effectiveness of the the TE model on these two tasks depends more heavily on the use of syntagmatic feature. Past research has found that syntagmatic associations often exist between words farther apart in text [Xu and Croft, 1996]. This may be the reason why the TE model performs better on these two tasks when a wider context window radius is

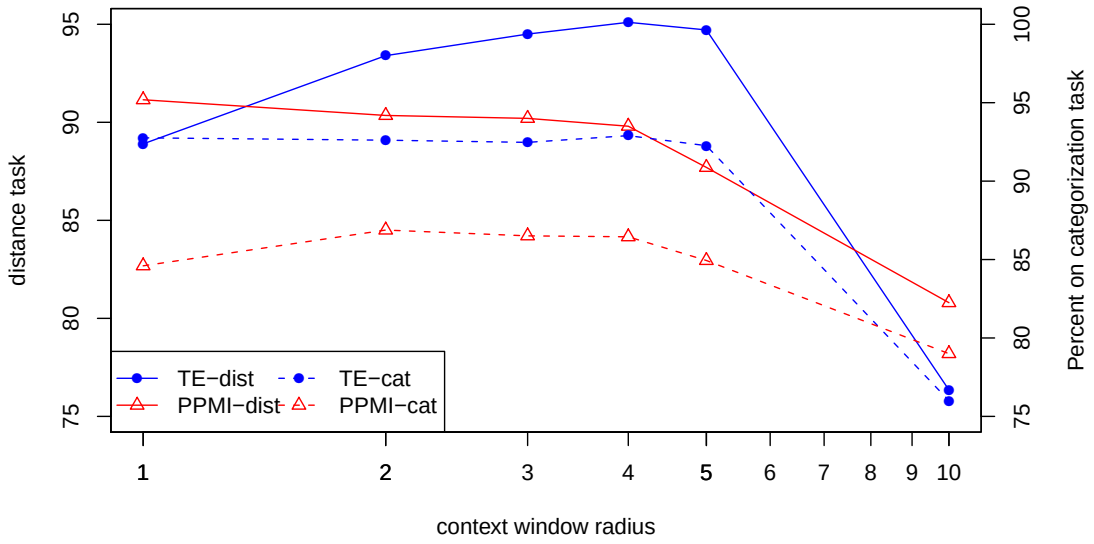

Figure 4.5: Sensitivity of the PPMI and TE models with respect to the context window radius for the semantic distance (dist) and semantic categorization (cat) tasks. The PPMI model was evaluated using vectors of 100,000 dimensions, while the TE model was evaluated with storage vector of 2,000 dimensions and a $\gamma = 0.1$.

used, when compared to that used when performing synonym judgements. However, given the performance of the TE model does not continue to increase past a context window radius of 5, the ability for the syntagmatic feature to excel on its own appears to be limited.

To investigate whether there are any computational costs associated with the TE model's superior task effectiveness, a complexity analysis of PPMI and TE is undertaken.

Computational Complexity Analysis

The number of dimensions used to store the geometric representations of terms within a semantic space impacts the models computational complexity. Like Bullinaria and Levy [2007] we chose to use context vectors created from the 100,000 most frequent vocabulary terms in our PPMI model. The storage complexity of the PPMI and TE models are proportional to the number of storage vectors $M(n) = |D_{\max}||V|$, where $|D_{\max}|$ is the dimensionality of the underlying storage vectors and $|V|$ is the size of the vocabulary. Therefore, PPMI has a storage complexity of $M(n) = 100,000|V|$. The time complexity of PPMI model is based on the operation used when calculating the cosine of the PPMI vectors, and is $T(n) = O(|D_{\max}|) =$

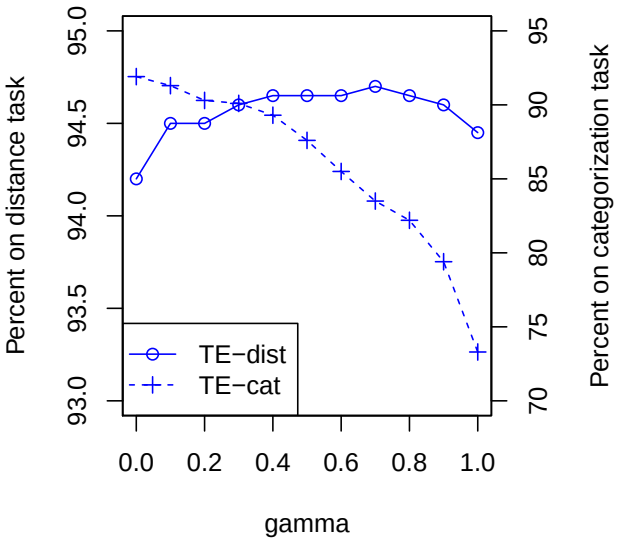

Figure 4.6: Sensitivity of the TE model with respect to gamma (γ) on the semantic distance (TE-dist) and semantic categorization (TE-cat) tasks, using a context window radius of 2 and storage vector of 500 dimensions.

$100,000$ basic operations.

The TE model stores the geometric representations in relatively small storage vectors. The impact of the dimensionality of the storage vectors (D_{SV}) on the TE model's performance for the semantic distance and categorization tasks is shown in Figure 4.7. These results demonstrate that the use of the TMC technique within the TE model provides robust performance on both tasks for $|D_{\text{SV}}| \geq 500$. Therefore, the storage complexity of the TE model using a storage vector of 500 dimension is $M(n) = 500|V|$. This is 200 times smaller than that of PPMI. The worst case time complexity of the TE model is determined from the time complexities of the syntagmatic and paradigmatic features outlined in Section 3.2.2, and is $T(n) = O(\frac{D_{SV}}{2}) + O(\frac{D_{SV}^2}{4})$. For $D_{SV} = 500$, this becomes $T(n) = \frac{500}{2} + \frac{500^2}{4} = 50,500$. Which is approximately half the time complexity of the PPMI approach.

This analysis suggests that the TE model is able to achieve superior effectiveness for less computational complexity when compared to PPMI on both tasks.

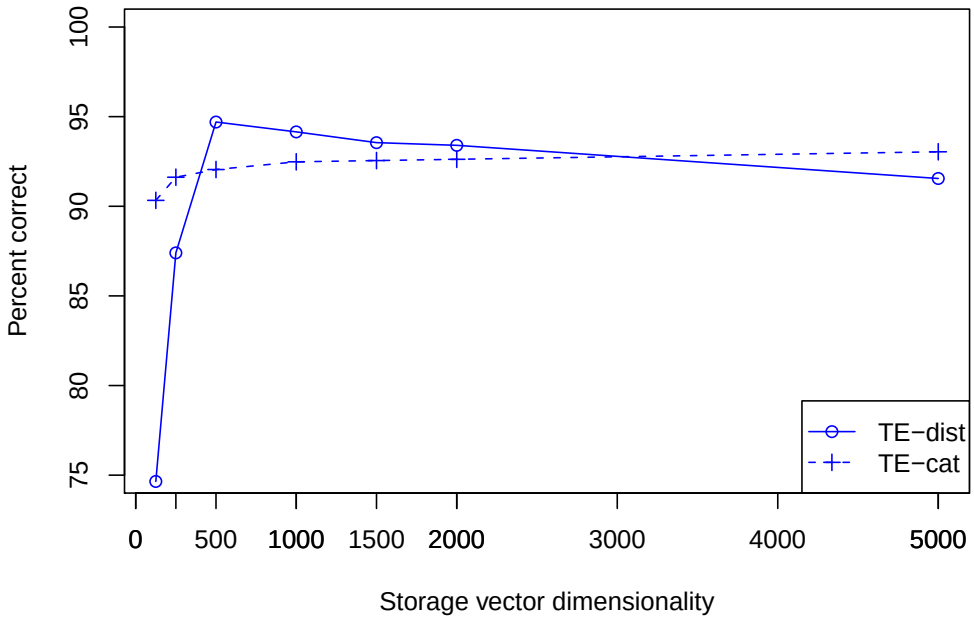

Figure 4.7: Sensitivity of the TE model with respect to storage vector dimensionality, using a context window radius of 2 and $\gamma = 0.4$ for the semantic distance (TE-dist) and semantic categorization (TE-cat) tasks.

4.3.3 Conclusion

The above evaluation of the TE model of word meaning, on a semantic distance and semantic categorization task, suggests that it provides superior task effectiveness when compared to a strong HAL based model (PPMI). Combined with the finding that the TE model achieves superior effectiveness over BEAGLE, PRI and PPMI on the TOEFL synonym judgement task, this result supports the conclusion that the TE model provides superior, robust performance on semantic tasks.

To continue to broaden the range of semantic tasks on which the TE model has been evaluated, it will now be used to judge the similarity of medical concepts on a number of benchmark data sets.

4.4 Similarity Judgement of Medical Concepts

The experiments outlined so far in this chapter relate to the domain of natural language, specifically tasks involving the modelling of semantic relationships between English words. These experiments have also used the same measures of syntagmatic and paradigmatic associations within the TE model's framework. To investigate how the TE model may perform on a more

specialised domain and to illustrate how different measures of syntagmatic and paradigmatic information can be used, a benchmark medical concept similarity task is evaluated.

4.4.1 Motivation

The ability to measure the similarity between medical concepts is important for overcoming issues faced by tasks involving medical documents, including medical search [Cohen and Widdows, 2009, Voorhees and Tong, 2011], literature-based discovery (e.g., drug discovery [Agarwal and Searls, 2009]) and clustering (e.g., gene clustering [Glenisson et al., 2003]).

As with natural language, determining the similarity between medical concepts faces challenges, including vocabulary mismatch. This can be seen by considering the medical terms *heart attack* and *myocardial infarction*, which refer to the same medical concept. If two different physicians use each term exclusively in their patient records, then when searching for patients who have experienced a *heart attack*, records relating to patients who have experienced a *myocardial infarction* should also be returned. To achieve some method of understanding the link between the two medical terms needs to be used.

Previous research into measuring the similarity between medical concepts compared path-based measures, which are based on the distances found between concepts in a medical thesaurus/ontology, with corpus-based approaches, like HAL and LSA [Pedersen et al., 2007]. This research highlighted the ability for corpus-based methods to provide superior task effectiveness.

A rigorous evaluation of eight different corpus-based approaches performing similarity judgements of medical concepts found a variant of the positive pointwise mutual information (PPMI) measure to be the most successful. This measure had a correlation of 0.66 with judgements made by expert human assessors [Koopman et al., 2012]. However, the ability for the PPMI measure to judge with such accuracy was not robust, and was shown to be sensitive to the training corpus.

4.4.2 Experimental Setup

Measures of Syntagmatic and Paradigmatic Associations

To evaluate whether the TE model can effectively perform similarity judgements of medical concepts a discussion of the most effective measures of syntagmatic and paradigmatic associations is required. Rather than re-invent the wheel, identifying which information the successful PPMI model is extracting may help guide the selection of syntagmatic and paradigmatic measures.

Similar to Equation (3.29), the PPMI measure of medical concepts captures the co-occurrence likelihood of concepts c_1 and c_2, and can be expressed as:

$$s_{\text{ppmi}}(c_1, c_2) = \begin{cases} \log\left(\frac{p(c_1,c_2)}{p(c_1)p(c_2)}\right) & \text{if } \log\left(\frac{p(c_1,c_2)}{p(c_1)p(c_2)}\right) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where $p(c_1, c_2)$ is the joint probability of c_1 and c_2, and $p(c_1)$, $p(c_2)$ are the expected probabilities of c_1 and c_2 respectively. However, the implementation of these estimates in Koopman et al. [2012], differed from those used in Bullinaria and Levy [2007] and our PPMI model used in the semantic judgement tasks carried out earlier in this chapter.

In practice, the probabilities underpinning the expressions in Equation (4.1) are computed as:

$$p(c_1, c_2) = \frac{|D_{c_1} \cap D_{c_2}|}{|D|} \quad p(c_1) = \frac{|D_{c_1}|}{|D|} \quad p(c_2) = \frac{|D_{c_2}|}{|D|}$$

where D_{c_2} is the set of documents containing concept c_2 and $|D|$ is the number of documents in the collection.

This implementation effectively models syntagmatic associations between concepts c_1 and c_2, except that the associations are modelled with a context window size equal to the document length, and the number of documents that the terms co-occur in is used to scale the score, rather than the document frequencies of each. Given PPMI's performance on this task, it was chosen as the syntagmatic measure to use within the TE model's framework, and will be referred to as **PPMI** in the following discussion.

For this task, the paradigmatic measure in Equation (3.30) was enhanced to boost scores for

concepts that display mainly paradigmatic associations, and is expressed as:

$$s_{\text{para}}(c_1, c_2) = \sum_{i \in V} \frac{f_{\overline{c_i c_1}} . f_{\overline{c_i c_2}}}{\max(\ f_{\overline{c_i c_1}}\ ,\ f_{\overline{c_i c_2}}\ ,\ f_{\overline{c_2 c_1}}\ )^2}, \tag{4.2}$$

where $f_{\overline{c_i c_1}}$ is the unordered co-occurrence frequency of concepts c_i and c_1, $f_{\overline{c_i c_2}}$ is the unordered co-occurrence frequency of concepts c_i and c_2, and V is the set of unique of concepts.

In this way, two concepts c_1 and c_2 that have strong syntagmatic associations; i.e., occur often within the same document and hence have a high $f_{\overline{c_2 c_1}}$ value in the denominator of Equation (4.2); will be penalised, resulting in a reduced paradigmatic score. The paradigmatic measure in Equation (4.2) will be referred to as **PARA** in the discussion below. For this experiment, the PARA measure was based on a semantic space built with a context window radius equal to the document length. Further explanation for this choice is provided when discussing the data sets in Section 4.4.2.

Given the choice of $s_{\text{ppmi}}(c_1, c_2)$ and $s_{\text{para}}(c_1, c_2)$ as the syntagmatic and paradigmatic measures, respectively, the semantic similarity of concepts c_2 given c_1 can be formally defined within the TE model as the probability of observing c_2 given c_1 in Equation (3.24), and expressed as:

$$\begin{aligned} P_{G,\Gamma}(c_2|c_1) &= \frac{1}{Z_\Gamma}\left[(1-\gamma)s_{\text{syn}}(c_1, c_2) + \gamma s_{\text{par}}(c_1, c_2)\right] \\ &= \frac{1}{Z_\Gamma}\left[(1-\gamma)s_{\text{ppmi}}(c_1, c_2) + \gamma s_{\text{para}}(c_1, c_2)\right] \end{aligned}$$

A more computationally efficient, rank equivalent expression becomes:

$$P(c_2|c_1) \propto (1-\gamma)s_{\text{ppmi}}(c_1, c_2) + \gamma s_{\text{para}}(c_1, c_2). \tag{4.3}$$

Data Sets

Two gold standard data sets were used to evaluate the effectiveness of the TE model on this task of judging the similarity of medical concepts. The first involves judging the similarity of 29 [7] *Unified Medical Language System* (UMLS) medical concept pairs. These pairs were first developed by Pedersen et al. [2007], and are provided with human assessments of semantic similarity produced by 9 clinical terminologists (coders) and 3 physicians. The correlation between groups of human assessors was $r = 0.85$. The method of human assessment required

[7] The pair *Lymphoid::hyperplasia* was removed from the original set of 30 as neither concept existed in the collections shown in Table 4.1

assessors to score each pair between 1 and 4, with 1 being unrelated and 4 being highly synonymous. This data set is referred to as the **Ped** data set and is provided in Appendix C.

The second data set, developed by Caviedes and Cimino [2004] is comprised of 45 UMLS concept pairs, for which semantic similarity assessments were performed by three physicians. Human assessments of each pair's similarity were scored between 1 and 10, with higher scores indicating a stronger similarity between concepts. This data set is referred to as the **Cav** data set and is provided in Appendix C.

To evaluate the sensitivity of the TE model to the training corpus, which was a weakness of the PPMI approach, two separate corpora were chosen for the evaluation. The first was the TREC MedTrack collection which consists of documents created from concatenating clinical patient records for a single visit. The second was the OHSUMED collection and is based on MEDLINE journal abstracts. The corpora statistics for each are shown in Table 4.1.

Corpus	# Docs	Avg. doc. len.	Vocab Size
TREC'11 MedTrack	17,198	5,010	54,546
OHSUMED	293,856	100	55,390

Table 4.1: Document collections (corpora) used for the medical concept similarity task.

As both the Ped and Cav data sets contained UMLS concept pairs the original text of the MedTrack and OHSUMED corpora needed to be converted to UMLS concepts. As was done by Koopman et al. [2012], the original textual documents were translated into UMLS medical concept identifiers using MetaMap, a biomedical concept identification system [Aronson and Lang, 2010]. This ensured that after processing the individual documents contained only UMLS concept ids. An example of the conversion process would be replacing the phrase *Congestive heart failure* in the original document with concept `C0018802` in the new document.

The implications of this conversion process on our measures can be seen by considering the paradigmatic measure. When modelling paradigmatic associations it is common to consider only those terms close to the target term, i.e., achieved by setting the context window radius within the TE model to one. However, given the conversion process and the fact that one term may be replaced by many concepts, or many terms replaced by one concept, the reasoning behind the use of narrow context windows to model paradigmatic associations may be invalid. Therefore, in this experiment the context window radius used to build the semantic space

underpinning the TE model's paradigmatic measure is set to the document length. This means that within the MedTrack corpus, paradigmatic associations are modelled using co-occurrences between concepts within an entire patient record, and in the OHSUMED corpus, paradigmatic associations are modelled using those found between concepts within an entire medical abstract.

4.4.3 Experimental Results

Given the TE model has a single parameter, a fair assessment requires the use of a training/test split of the data sets chosen. This is achieved by training γ on one data set and then using the tuned γ value to test on a separate data set. The train/test split chosen for this experiment is shown in Table 4.2 along with the correlation scores with human judgement (denoted by r) achieved by the TE and PPMI models. These results have also been reported in Symonds et al. [2012c].

Test: Corpus (data set)	Training: Corpus (data set)	γ	TE	PPMI
MedTrack (Ped)	OHSUMED (Ped)	0.5	$r = 0.6706$	$r = 0.4674$
MedTrack (Cav)	MedTrack (Ped)	0.5	$r = 0.6857$	$r = 0.6154$
OHSUMED (Ped)	OHSUMED (Cav)	0.2	$r = 0.7698$	$r = 0.7427$
OHSUMED (Cav)	MedTrack (Cav)	0.4	$r = 0.8297$	$r = 0.8242$

Table 4.2: Performance of the TE model using the γ produced by the specified train/test splits; performance of PPMI included for comparison. The average across all data sets are 0.74, 0.66 for the TE and PPMI approaches, respectively.

A graphical comparison of the ability of the TE, PPMI and PARA approaches to correlate with expert human assessors on the gold standard data sets is shown in Figure 4.8. This graph suggests that the TE model achieves a much higher correlation with human judged similarity scores (with an average correlation of 0.74 over all datasets and corpora) compared to the paradigmatic (PARA: 0.57) and syntagmatic (PPMI: 0.66) measures on their own.

Even though the TE model achieves an average of 12% improvement over the PPMI approach, a paired t-test on the 4 unique combinations of data sets and training corpus does not show this improvement to be statistically significant at the 95% confidence interval ($\alpha = 0.05$). Given there are only 4 samples this result is not surprising. However, future work to test the

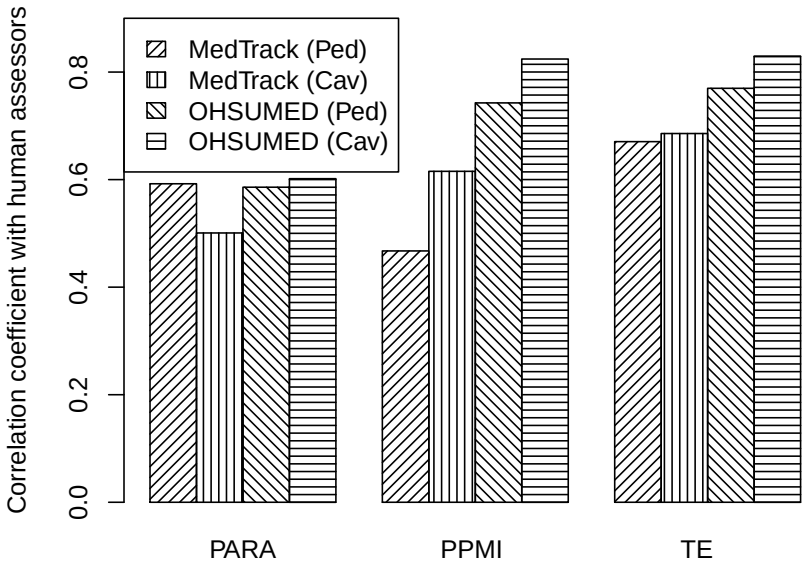

Figure 4.8: Correlation coefficients for medical concept similarity judgements produced by PARA, PPMI and TE with those provided by human expert assessors.

significance of the TE model's improvements on this task could be undertaken by evaluating it on other data sets and training corpora.

To gain a broader understanding of how the TE model's effectiveness on this task compares to the many popular corpus-based approaches, Figure 4.9 displays the average effectiveness of the TE model and PARA measure in relation to 14 other corpus-based approaches reported in Koopman et al. [2012], on the data sets and document collections used in this experiment[8]. This comparison is provided to show the range of effectiveness of existing corpus-based methods, and to illustrate that the improvement of the TE model over PPMI is above the average trend of the overall graph.

Parameter Sensitivity

An investigation into the effectiveness achieved for various mixes of syntagmatic and paradigmatic information can provide a better understanding of the importance of each type of information on this task. The sensitivity of task effectiveness to the TE model's γ parameter, which is responsible for this mix, is shown in Figure 4.10. This graph shows that for all data set and

[8]For further details on the settings used within the 14 corpus-based approaches evaluated in Koopman et al. [2012], the reader is referred to the paper

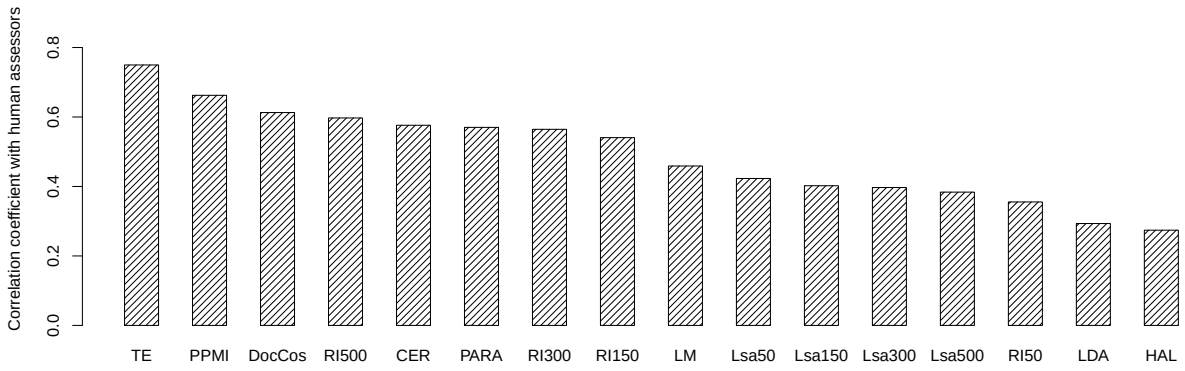

Figure 4.9: Comparison of effectiveness of a HAL based model (HAL), Latent Dirichlet Allocation (LDA) model, RI model using 50, 150, 300 and 500 dimensions, LSA model using 50, 150, 300 and 500 dimensions, Document Vector Cosine (DocCosine) model, Cross Entropy Reduction (CER) model, a language model (LM), PPMI and our TE and PARA implementations. Correlations scores are averaged across datasets and training corpora. For further details on the settings used within the HAL, LDA, LSA, RI, DocCosine, CER and LM models refer to Koopman et al. [2012]

training corpus combinations the best correlation with human expert assessors is achieved when both syntagmatic (PPMI) and paradigmatic (PARA) information is used by the TE model. It also demonstrates the increased robustness of the TE model over the PPMI approach.

The robustness of the PPMI and PAR measures across datasets and corpora can be seen by comparing the distance between the *end points* of the lines drawn in Figure 4.10. The left hand side of the graph (where $\gamma = 0$) illustrates the performance of the TE model when only syntagmatic information is modelled, i.e. when the estimate of similarity becomes solely based on the PPMI measure (as Equation (4.3) reduces to $s_{\text{ppmi}}()$ when $\gamma = 0$).

The right hand side of the graph ($\gamma = 1$) shows the performance of the TE model when only paradigmatic information is used, i.e. when the estimate of similarity is solely based on the PARA measure. With most lines converging to the same point on the right hand side, this demonstrates the increased robustness information about paradigmatic associations can provide. Therefore, the TE model allows the robustness of the PARA measure to be combined with the strong average effectiveness of the PPMI measure across a wide range of training corpora. To gain further insight into how this is being achieved, a more detailed analysis is required.

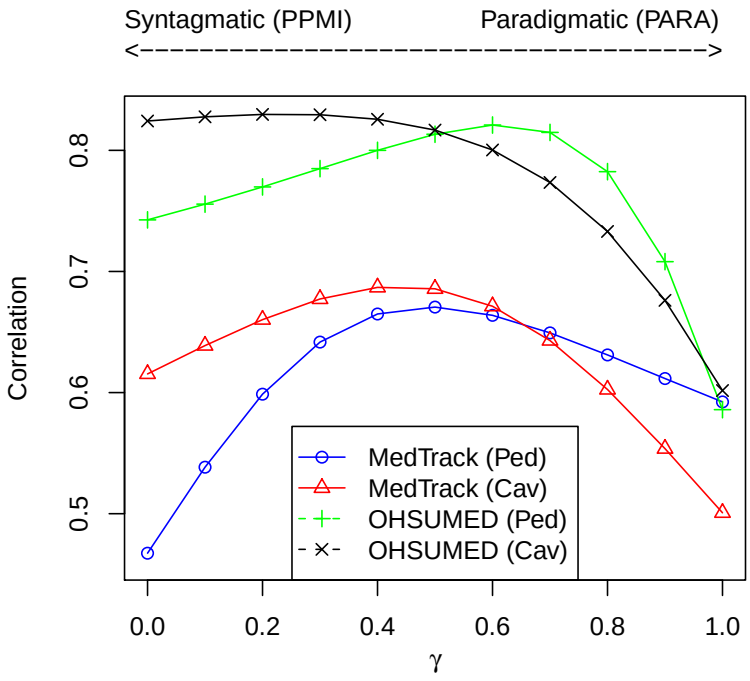

Figure 4.10: Sensitivity of the TE model's ability to match human expert assessors judgements of medical similarity for various values of γ.

Analysis of Paradigmatic and Syntagmatic Behaviour

To illustrate in more detail why the combination of both the paradigmatic and syntagmatic measures can achieve such strong and robust effectiveness across all datasets and corpora, the correlation of PPMI, PAR and TE is compared to that of expert human assessors on a per concept-pair basis.

Figure 4.11 illustrates the log-based normalised similarity scores of human assessors, PPMI, PARA and TE for the Caviedes and Cimino (Cav) dataset when PPMI, PARA and TE are trained on the OHSUMED corpora. The concept-pairs are placed in descending order of similarity as assessed by human judges, i.e. from the most similar human judged pairs to the least from left to right. The performance of a measure can be visualised by comparing the trend of its graph with that of the descending *human assessed* graph. If a measure's graph trends in a similar fashion to that of the human assessed graph then it is considered to have a stronger correlation with similarity judgements made by human expert assessors.

To better understand why the paradigmatic based measure (PARA) differs from human assessors in Figure 4.11, the document frequencies of concept pairs 11, 16, 21, 34, 36, 37

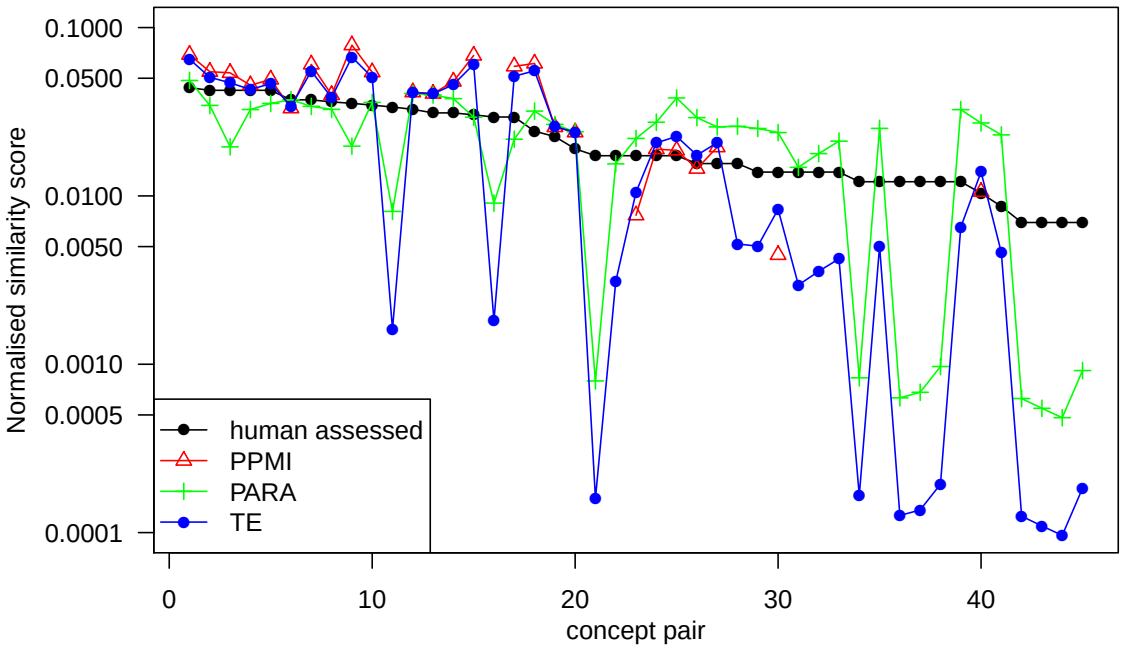

Figure 4.11: Normalised similarity scores (on log scale) of human assessors, PPMI, PARA and TE on Caviedes and Cimino dataset (Cav) when using the OHSUMED corpus for training.

and 38 from the Cav data set are reported in Table 4.3.

Table 4.3 shows that for these concept pairs at least one concept occurs in a very small number of documents. This provides little evidence for the accurate estimation of paradigmatic associations between the concept pairs. We therefore conclude that the PARA measure requires concepts to occur in a sufficient number of documents for an effective similarity estimate to be made.

Similar observations are valid across datasets and corpora. For example, consider the correlation of PARA with human judgements for the Pedersen et al. (Ped) data set and the MedTrack corpus, as shown in Figure 4.12. The document frequencies for a number of concept pairs that show divergence from the Ped data set are shown in Table 4.4.

For these concept pairs where the estimate of similarity made by the PARA measure diverges from human judges, the PPMI measure more effectively estimates the similarity. Thus the TE model, which mixes the two forms of associations, is still effective even when the PARA measure is unreliable. This further supports the inclusion of both paradigmatic and syntagmatic associations for assessing semantic similarity between medical concepts.

Pair #	Concept 1	Doc. Freq.	Concept 2	Doc. Freq.
11	Arrhythmia	2,298	Cardiomyopathy, Alcoholic	13
16	Angina Pectoris	1,725	Cardiomyopathy, Alcoholic	13
21	Abdominal pain	690	Respiratory System Abnormalities	1
34	Cardiomyopathy, Alcoholic	13	Respiratory System Abnormalities	1
36	Heart Diseases	1,872	Respiratory System Abnormalities	1
37	Heart Failure, Congestive	1,192	Respiratory System Abnormalities	1
38	Heartburn	104	Respiratory System Abnormalities	1

Table 4.3: Example Cav concept pairs for which the PARA measure diverges from the human judgements on the OHSUMED corpus. Document frequencies showing the prevalence of the concepts in the corpus are reported.

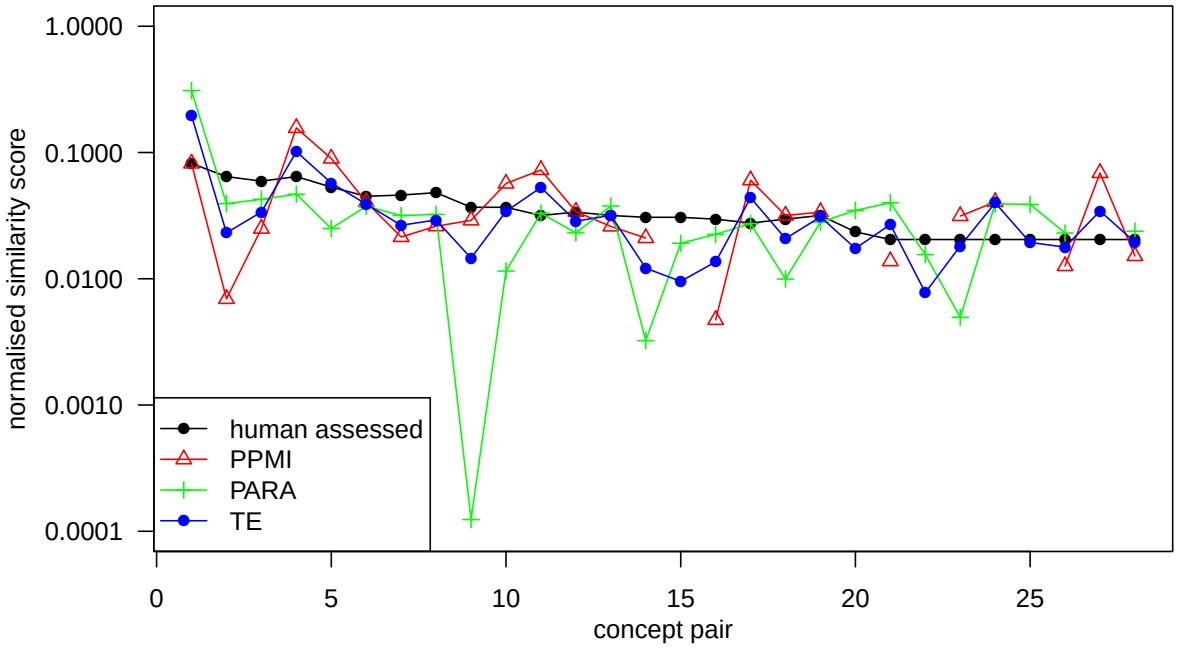

Figure 4.12: Normalised similarity scores (on log scale) of human assessors, PPMI, PARA and TE on the Pedersen et al. dataset (Ped) when using the MedTrack corpus for training.

Pair #	Concept 1	Doc. Freq.	Concept 2	Doc. Freq.
9	Diarrhea	6,184	Stomach cramps	14
23	Rectal polyp	26	Aorta	3,555

Table 4.4: Example Ped concept pairs for which the PARA measure diverges from the human judgements on the MedTrack corpus. Document frequencies showing the prevalence of the concepts in the corpus are reported.

Figure 4.11 also illustrates a number of *discontinuities* in the PPMI graph. A discontinuity, i.e. the absence of the data-point within the plot, is due to a PPMI score of zero for the concept pair[9]. In practice, these discontinuities represent instances where the concepts in the pair never co-occur within any document. The same situation applies across other datasets and corpora, for example the Ped data set on MedTrack corpus shown in Figure 4.12.

While PPMI discontinuities for concept pairs judged as unrelated by human assessors are correct estimates (as PPMI$= 0$ implies they are unrelated), discontinuities for concept pairs judged similar by human assessors(e.g. pairs 11, 16, etc. in Figure 4.11) indicate a failure of the PPMI measure. These latter examples, where no within document co-occurrences exist between concepts being considered, highlight a likely reason why the effectiveness of approaches that rely solely on *syntagmatic* information to judge similarity, such as PPMI, are sensitive to the choice of training corpus.

However, given the superior effectiveness achieved by the TE model on all data sets, we can conclude that appropriately mixing both syntagmatic and paradigmatic associations can help address these training corpus sensitivity issues.

4.4.4 Conclusion

This evaluation has demonstrated the robustness and effectiveness of the TE model in judging the similarity of medical concepts, with respect to their correlation with expert human assessors. By explicitly modelling syntagmatic and paradigmatic associations the TE model is able to outperform state of the art corpus-based approaches. Furthermore, the TE model is robust across corpora and datasets, in particular overcoming training corpus sensitivity issues experienced

[9]As the graph is in log scale, $\log(0) = -\infty$ cannot be plotted.

by previous approaches. This result appears to be due to the use of information about both syntagmatic and paradigmatic associations between medical concepts, and most importantly due to the TE model's ability to control the combination of these two sources of information.

A possible area for future work is the development of an adaptive TE approach. An adaptive approach would determine the best mix of syntagmatic and paradigmatic information on a per concept-pair basis, using corpus statistics. This analysis has shown that paradigmatic associations require a minimum number of *occurrences of concepts* within the corpus. While, syntagmatic associations require a minimum number of *co-occurrences of concept pairs* within documents. These corpus statistics could represent features for a machine learning approach to predict the optimal mix of syntagmatic and paradigmatic information for a given concept pair.

4.5 Summary

This chapter has evaluated the performance of the TE model, developed in Chapter 3, on a wide variety of semantic tasks, including synonym judgement, semantic distance, semantic categorization and similarity judgements of medical concepts. The superior effectiveness of the TE model on these tasks when compared to strong benchmark approaches demonstrates the versatility and effectiveness of the TE model's formal framework.

The computational complexity analyses of the TE model and other current corpus-based SSMs, outlined in Chapter 3 and in Section 4.3.2, highlighted the efficiency gains provided by the novel TMC technique, developed to store the sparse tensor representations formed within the TE model's semantic space.

This concludes the theory development of the TE model of meaning. Part II of this dissertation will demonstrate how the TE model can be applied within the area of information retrieval, along with a rigorous evaluation of its performance.

Part II

Application to Information Retrieval

Chapter 5

Information Retrieval

5.1 Overview

The story so far has highlighted the belief that semantic technologies, particularly those using SSMs, may transform communication between humans and computers (Section 1.4). A review of advances in SSM development drew out a number of key enhancements that appear likely to underpin the next generation of SSMs (Chapter 2). These ideas were used to develop a novel SSM that efficiently captured information about syntagmatic and paradigmatic associations between words so that these linguistic associations could be combined within a formal model of word meaning (Chapter 3). This tensor encoding (TE) model of word meaning was then evaluated against a number of benchmark models and demonstrated superior performance across a wide range of semantic tasks (Chapter 4).

In Section 1.4, a hypothesis that significant improvements in retrieval effectiveness could be achieved if a model of word meaning, accessing both syntagmatic and paradigmatic associations, was used to augment query representations within the information retrieval process. This was due to the heavy dependence on word meanings when a user formulates their query and the lack of paradigmatic information used in existing query expansion techniques.

To test this hypothesis an overview of a number of relevant concepts in information retrieval is required (Chapter 5), along with the development of a formal query expansion technique that incorporates the TE model (Chapter 6) and finally a rigorous evaluation of this technique (Chapter 7).

This chapter provides an introduction to a number of key concepts within the field of

information retrieval, most importantly *relevance*. A brief overview of classic probabilistic models, on which most state-of-the-art models are based, leads to a discussion on augmenting query representations within these models. It is within this query augmentation process that the TE model will be applied. Therefore, justification for the choice of appropriate benchmark models, against which to evaluate the TE model's performance, is provided here.

5.2 An Introduction to Information Retrieval

Information retrieval is a widely, often loosely-defined term. This is because it spans many areas, including document retrieval, question and answering tasks, and library systems, and web search, to name a few. However, generally an automated information retrieval system includes the automated process of assessing the *relevancy* of information (not data) for a given search criterion, commonly between a query and a set of documents, as shown in Figure 5.1.

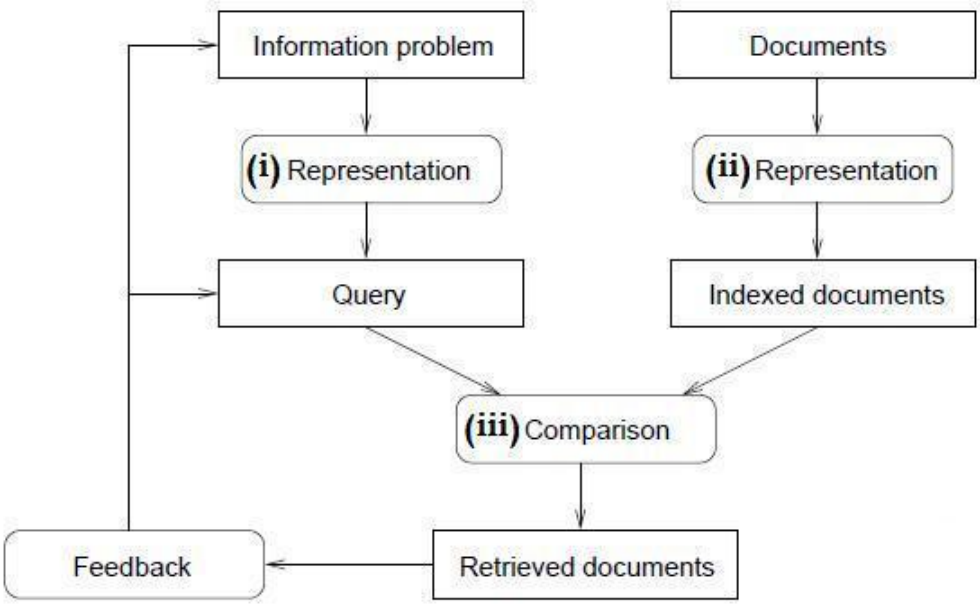

Figure 5.1: The document retrieval process (adapted from [Croft et al., 2009]).

The information retrieval process is inexact, inductive and probabilistic; as opposed to data retrieval which is normally exact, deductive and deterministic [van Rijsbergen, 1979]. The inaccuracy of the information retrieval process stems largely from the uncertainty in evaluating *relevance*.

In a detailed summary of how the concept of relevance in information retrieval has been

represented over the years, Mizzaro [1998] outlines four dimensions that allow the types of relevance to be classified with respect to *the relevance interesting for the user*:

1. Information resources: {Surrogate, Document, Information}
2. Representation of the user's problem: {Real Information Need (RIN), Perceived Information Need (PIN), Request, Query}
3. Time: $\{t_0, \ldots, t_f\}$
4. Components: {Topic, Task, Context}

Mizzaro's second dimension, *representation of the user's problem*, captures the difficulties of natural language to express a user's information need. Acceptance that a user's query is an imprecise description of their information need has appeared since the earliest research into information retrieval [Cleverdon et al., 1966]. This is even more pertinent when considering the short, two or three word queries submitted to modern web search engines [Carpineto and Romano, 2012]. For these reasons there is a strong interest in the use of query expansion techniques to augment the query to be a more precise representation of the information need. Such techniques have been shown to significantly increase average retrieval effectiveness [Lavrenko and Croft, 2001, Lv and Zhai, 2010, Metzler and Croft, 2007, Zhai and Lafferty, 2001]. However, this effectiveness is known to be sensitive to parameter selection and often deteriorates performance for some queries [Billerbeck and Zobel, 2004, Collins-Thompson, 2009].

Many state-of-the-art query expansion techniques do not explicitly model the associations that exist between words in natural language. However, some of the approaches developed in the last decade have demonstrated that even greater improvements in retrieval effectiveness can be achieved by incorporating word (or term) dependency information [Lv and Zhai, 2010, Metzler and Croft, 2007]. Many of these term dependency based approaches use the intuition that useful expansion terms co-occur in context with the query terms within a document [Metzler and Croft, 2007, Xu and Croft, 1996].

It is interesting to note that this intuition is the same as that behind the notion of *syntagmatic associations*. However, as highlighted by structural linguistic theory, paradigmatic associations are also important when modelling the word meanings that are likely drawn upon when a user formulates their query based on their information need. The TE model provides a novel way for

information about these word meanings to be accessed, by explicitly modelling and combining information about syntagmatic and paradigmatic information. This could be used to enhance query representations within the information retrieval process. However, to determine how best to use the TE model to expand query representation an understanding of existing information retrieval models and how they are evaluated is required.

The following sub-sections outline relevant background on the information retrieval process, including how the effectiveness of existing models are evaluated. They second section of this chapter will provide an overview of current query expansion techniques, including their efforts to model information about the dependencies that naturally exist between words in natural language.

5.2.1 Evaluating Information Retrieval Systems

The evaluation methods for determining the effectiveness of information retrieval systems are based on the concepts of *Precision* and *Recall*, first used by Cleverdon [1970] and Salton [1968].

Precision is defined by the ratio of *the number of relevant documents retrieved* to *the total number of documents retrieved*, and can be stated as:

$$\text{precision} = \frac{|Rel \cap Ret|}{|Ret|}, \tag{5.1}$$

where Rel represents the set of *relevant* documents (both retrieved and not retrieved) in the system, and Ret represents the set of documents currently returned.

Precision measures the amount of noise being produced by an information retrieval system. If all returned documents are also relevant, then $precision = 1$ and there is considered to be no noise generated from the system.

Recall is defined by the ratio of *the number of relevant documents retrieved* to *the total number of relevant documents (both retrieved and not retrieved) in the system*, and can be stated as:

$$\text{recall} = \frac{|Rel \cap Ret|}{|Rel|}. \tag{5.2}$$

Recall is used to measure the level of omission or loss of the system. A low recall indicates a lot of relevant information is being omitted. Once all relevant documents have been returned, recall $= 1$.

The *effectiveness* of an information retrieval system can be measured by using both recall and precision. Ideally a system will attempt to maximize recall and precision, however it has been shown experimentally that these two measures interact with each other, such that a trade-off between recall and precision is often required. Depending on the purpose of the information retrieval application one may prefer recall over precision, or *vice versa*. For example, for litigation lawyers looking for precedents, the ability to have all relevant documents returned (i.e., recall $= 1$) is of greater value than a high precision.

The other important measure of performance used in information retrieval system development is *efficiency*. Efficiency relates to the amount of computing memory and CPU time required to perform the information retrieval task. Improving both effectiveness and efficiency is the aim of information retrieval research and development [van Rijsbergen, 1979].

Evaluation of retrieval performance often requires large scale testing platforms on which to simulate web scale environments. These evaluation platforms are still based on the ideas introduced in the Cranfield experiments [Cleverdon et al., 1966], where large document sets are combined with queries and relevance judgements to allow automated testing and calculation of recall and precision. Relevance judgements are made up of a list of documents indicating their relevance, which can be binary (i.e., relevant or non-relevant), or graded (i.e., using a score between 0 and 4, with 0 being not relevant and 4 being extremely relevant, say).

Due to the large collection sizes used in modern information retrieval evaluation, the determination of relevance judgements for queries are normally facilitated through the use of pooling methods. Pooling involves judging only the top ranked documents returned by a number of retrieval models. The *Text REtrieval Conference* (TREC), which is designed to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies, has been used to develop a number of benchmark data sets on which to evaluate systems. Standard TREC data sets (including documents, queries and relevance judgements) will be used to evaluate the use of the TE model within the query expansion process (Chapter 7). TREC also provides an opportunity for the industry and researchers to compete on a number of information retrieval tasks. The approach developed in this research was also be submitted to the TREC Web track for evaluation against other state-of-the-art systems.

Measures of Effectiveness in Ad Hoc Retrieval

Various measures of retrieval effectiveness have been developed and used in the past. These can be grouped into two major classes based on the type of relevance judgements they rely on, which are: (i) binary measures and (ii) graded measures. Binary measures rely solely on a judgement of relevance or non-relevance for each document. However, graded measures rely on estimates of graded relevance for each document, such as non-relevant, slightly relevant, mostly relevant or extremely relevant. The following section outlines two measures from each class that are used to evaluate TQE within this research.

Binary Measures: Precision and Recall underpin many popular measures of retrieval effectiveness. Two popular binary measures, i.e., based on relevance judgements of relevant or non-relevant, include *mean average precision* (MAP) and *precision at* k (P@k). The MAP for a set of queries is the mean of the average precision scores for each query, where the average precision is measured by summing the precision and recall scores at each position in the retrieved document list:

$$AveP = \sum_{k=1}^{n} P(k)\Delta r(k) \tag{5.3}$$

where k is the rank in the sequence of retrieved documents, n is the number of retrieved documents, $P(k)$ is the precision at cut-off k, and $\Delta r(k)$ is the change in recall from $k-1$ to k.

The precision at k (P@k) measure, is computed as the precision score, refer to Equation (5.1), after the top $|Rel| = k$ retrieved documents. Both MAP and P@20 have a long track record of being employed to measure information retrieval effectiveness [Croft et al., 2009].

Graded Measures: Graded metrics aim to enhance the effectiveness score for systems that return relevant documents towards the *head* of the ranked list. In effect discounting the importance of relevant documents ranked lower. In contrast P@20 has no discounting, as it is purely a percentage calculation and MAP exhibits a linear discounting. Graded metrics have been used heavily in evaluating the effectiveness of systems performing web search, and include (*expected reciprocal rank* (ERR) at k, *normalised discounted cumulative gain* (nDCG) at k).

ERR measures the (inverse) expected effort required for a user to satisfy their information

need, and is defined at rank k as:

$$ERR@k = \sum_{i=1}^{k} \frac{R(g_i)}{i} \prod_{j=1}^{i-1} (1 - R(g_i), \tag{5.4}$$

where k is the number of results being considered ($n = 20$ in this experiment), and $R(g_i)$ is the probability that the user is satisfied with result i, which is computed as:

$$R_i = \frac{2^g - 1}{2^{g_{\text{max}}}}, \tag{5.5}$$

where g is the editorial grade associated with the relevance judgement of that document (e.g., $g \in [0, 4]$ say, where $g = 0$ indicates a document that is non relevant and $g = 4$ indicates a document considered highly relevant to the query).

The *normalised discount cumulative gain* (nDCG) measure aims to penalise the existence of highly relevant documents being lower down in the search results. The measure calculated at a particular rank position k can be computed as:

$$nDCG@k = \frac{DCG@k}{IDCG@k}, \tag{5.6}$$

where $IDCG$ is the ideal DCG computed for the top k documents being order from most relevant to least, and

$$DCG@k = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{\log_2(1 + i)}. \tag{5.7}$$

The key difference between ERR and nDCG is that ERR heavily discounts the contributions of documents that appear after highly relevant documents.

5.2.2 Document Retrieval Models

Information retrieval models can be classified into three broad categories based on the mathematical framework within which they are positioned, including (i) *logical* models (based on mathematical logic, including set theory), (ii) *vector space* models (based on geometry and linear algebra) and (iii) *probabilistic* models (based on probability theory). Most state-of-the-art information retrieval models are probabilistic in nature. Therefore, using the TE model to augment query representations within probabilistic retrieval models is argued to be a sound initial setting within which to evaluate its use.

Classical Probabilistic Models

Maron and Kuhns [1960] were the earliest researchers to develop the link between the critical notion of *relevance* in information retrieval and the comparative concept of relevance explicated in the theory of probability. However, it was many years later before any significant headway was made on probabilistic methods [van Rijsbergen, 1979].

A formal framework, known as the Probability Ranking Principle (PRP) [Robertson, 1977] was developed to justify optimal ranking of documents, and states:

> "The Probability Ranking Principle (**PRP**): If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data."

This principle relies on a number of assumption, most significantly for this research, an assumption of information independence that weakens, if not removes, any acknowledgement that dependencies exist between the words found in documents. A family of information retrieval models are based on this principle, known as classical probabilistic models, and include:

- **The Binary Independence Retrieval (BIR) model**: As the name may suggest the BIR model [Robertson and Sparck Jones, 1988] uses a binary representation of terms within documents and assumes term independence. The importance of within-document term frequencies in the estimating relevance of a document meant that the binary nature of the model was a major drawback [Luhn, 1958, Robertson, 1981].

- **2-Poisson model**: Statistical work by Harter [1975] found that specialty words (containing most of the information content) occurred more frequently in *elite* documents, whereas non-specialty words (like those found in stop lists) were randomly distributed over the collection. This led to the development of the 2-Poisson model, which was not able to be practically implemented due to the number of hidden parameters, but guided researchers in the use of term frequencies in calculating probabilities and led to the BM25 model.

- **BM25** (Best Match 25): Inspired by the 2-Poisson model, the BM25 model [Robertson and Walker, 1994] has proven to be one of the most successful information retrieval models and has a ranking function commonly reported as:

$$P(R=1|D) \sim \sum_{j \in Q \cap D} tf_{j,Q} \frac{(k_1+1)tf_{j,D}}{k_1\left((1-b)+b\frac{|D|}{|D|_{avg}}\right)+tf_{j,D}} log\left(\frac{C-df_j+0.5}{df_j+0.5}\right), \tag{5.8}$$

  where $tf_{j,Q}$ is the number of times term j appears in the query Q, $tf_{j,D}$ is the number of times query term j occurs in document D, $|D|$ is the length of document D, $|D|_{avg}$ is the average document length of the collection, C is the number of documents in the collection, df_j is the number of documents term j occurs in, and k_1 and b are tuning parameters for the model. This expression ignores term dependencies, as it accumulates scores based on individual terms (j) and there is no provision for expressing the dependencies between terms. The first part of the ranking expression (involving parameters k_1 and b) is considered the term frequency like component (indicating aboutness) and the second part an inverse document frequency (idf) like component, indicating informativeness [Jones, 2004].

- **Divergence From Randomness (DFR) model**: Based on Harter's notion of eliteness, the DFR model [Amati and Van Rijsbergen, 2002] identifies significant terms as those that diverge the most from the random background model. The DFR model presents a framework for developing various term weighting schemes for use in a ranking function similar to the BM25 in structure. The difference being that the tf and idf components are based on various randomness models, including Binomial and Bose-Einstein distributions (geometric).

Unigram Language Models

Due to the dominance of heuristic models, like the BM25, other forms of probabilistic models were pursued. A formal model is preferred as it provides more control over the assumptions associated with a model, which allows greater understanding of the implications of extending the model. The use of statistical language modeling in information retrieval [Hiemstra, 2001, Ponte and Croft, 1998] was adapted from the field of artificial intelligence, where it had been successfully used for applications including speech recognition [Charniak, 1994]. In a statistical language model the key elements are the probabilities of word sequences, denoted

as $P(w_1, w_2, \ldots, w_n)$ or $P(w_1, n)$ for short. Normally, due to the length of sentences and the cost of estimating long word sequences, the statistical language model is approximated by the following n-gram models

- Unigram: $P(w_1, n) = P(w_1)P(w_2)\ldots P(w_n)$
- Bigram: $P(w_1, n) = P(w_1)P(w_2|w_1)\ldots P(w_n|w_{n-1})$
- Trigram: $P(w_1, n) = P(w_1)P(w_2|w_1)P(w_3|w_{1,2})\ldots P(w_n|w_{n-2,n-1})$

The *unigram* language model is the simplest and most popular form of the language model in information retrieval and applies a probability distribution over the words based on the assumption of word independence.

One of the most common strategies for using language models for information retrieval is the *query likelihood* approach, which treats documents as language models and queries as a generation process. In this case a document is considered relevant to a query based on the probability that the query can be generated from the document $P(Q|D)$. This assumption that relevance is linked to query generation from a document differs from the Probability Ranking Principle (PRP), which underpins classical probabilistic models. This significant step removes the explicit relevance variable R. Thus, starting with a query $Q = (q_1, \ldots, q_p)$ and document D as evidence, documents are ranked according to the likelihood $P(D|Q)$ that the query was generated. Using Bayes Rule the ranking becomes:

$$P(D|Q) \propto P(Q|D)P(D), \tag{5.9}$$

where $P(D)$ is the prior probability of D and is often considered equal for all documents, i.e., uniform. Therefore, when $P(D)$ is removed and $P(Q|D) = \prod_{i=1}^{p} P(q_i|D)$ substituted to show the probability of each query term multiplied together, a rank equivalent form of Equation (5.9) becomes:

$$P(D|Q) \propto \prod_{i=1}^{p} P(q_i|D). \tag{5.10}$$

As the probabilities ($P(q_i|D)$) involved are normally small, the logarithm of both sides of Equation (5.10) is often taken, to provide a final rank equivalent form:

$$P(D|Q) \propto \sum_{i=1}^{p} \log\left(P(q_i|D)\right). \tag{5.11}$$

To calculate the right hand side of Equation (5.11), an estimate for $P(q_i|D)$ is required. A first approach would be to base the estimate on the relative frequencies of the query terms within the document:

$$P(q_i|D) = \frac{tf_{q_i,D}}{|D|}, \tag{5.12}$$

where $tf_{q_i,D}$ is the frequency of a query term q_i in document D, $|D|$ is the length of the document. This approach is known as maximum likelihood estimation. Note how the independence assumption between query terms carries across into computations involving the document. The main issue with this estimation is that if a document did not contain all of the query terms it would have a ranking score ($P(D|Q)$) of zero, i.e., since the joint probability is a multiplication of probabilities. To reduce this issue a form of smoothing is introduced. Smoothing normally involves discounting the probability estimate of the missing terms.

Several forms of smoothing have been proposed [Zhai and Lafferty, 2004], these include *Jelinek-Mercer* (JM) smoothing which uses linear interpolation and is expressed as:

$$P(q_i|D) = (1-\lambda)\frac{tf_{q_i,D}}{|D|} + \lambda\frac{cf_{q_i}}{|C|}, \tag{5.13}$$

where cf_{q_i} is the frequency of term q_i in the collection C, and $|C|$ is the length of the collection.

Substituting Equation (5.13) into Equation (5.11), the final JM smoothed ranking function becomes:

$$P(D|Q) \propto \sum_{i=1}^{p} log\left((1-\lambda)\frac{tf_{q_i,D}}{|D|} + \lambda\frac{cf_{q_i}}{|C|}\right). \tag{5.14}$$

Another successful and popular smoothing approach is the Dirichlet smoothing method, which uses the following estimate:

$$P(q_i|D) = \frac{tf_{q_i,D} + \mu\frac{cf_w}{|C|}}{|D| + \mu}, \tag{5.15}$$

where μ is the Dirichlet smoothing parameter.

Language modelling has become a popular, robust and highly effective approach to building retrieval models. Its performance has been demonstrated to be equivalent to BM25 with the additional benefit that it is formally developed.

Dependency Based Models

The document ranking models outlined so far *do not* explicitly consider information about the dependencies between words found in natural language. Although researchers have argued

for their inclusion in the document ranking process [van Rijsbergen, 1977], early efforts to include term dependencies within document retrieval models based on phrases have met with mixed success [Croft et al., 1991]. However, more recent efforts, using statistical information based on co-occurrence patterns of words have shown statistically significant improvements in retrieval effectiveness, including work done by Metzler and Croft using the Markov random field model [Metzler and Croft, 2005] and research by Tao and Zhai incorporating proximity information into a number of existing retrieval models [Tao and Zhai, 2007].

Markov Random Field

The *Markov random field* (MRF) model [Metzler and Croft, 2005] provides a formal method for combining various statistical features of the underlying document collection. Metzler and Croft [2005] reported that past dependency models have failed in achieving robust *state-of-the-art* performance due to data sufficiency reasons, including:

1. Past models have been based on the Binary Independence Retrieval (BIR) model. Therefore the term dependencies must be estimated in both relevant and non-relevant classes, where there is often very little sample data to make these estimates.

2. Smaller document collections used by past researchers testing term dependency models have consisted of very short documents from which they believe there is very little hope of accurately modeling term dependencies.

For these reasons, Metzler [2007] argues that past dependency models have themselves not been flawed.

Metzler [2007] also states that as corpus sizes increase the ability for inverse document frequencies and term frequency weightings to provide robust performance will diminish due to the increasing noise. He believes the key will be in the ability for term dependency models to provide a more powerful discriminating role in these noisy systems.

As the name suggests, the MRF model uses a Markov random field to formally combine various features used in estimating the probability of observing a document D, given a query Q. As outlined in Section 3.2.1, a Markov random field is an undirected graph (e.g., Figure 5.2) that can be used to estimate a conditional probability between variables.

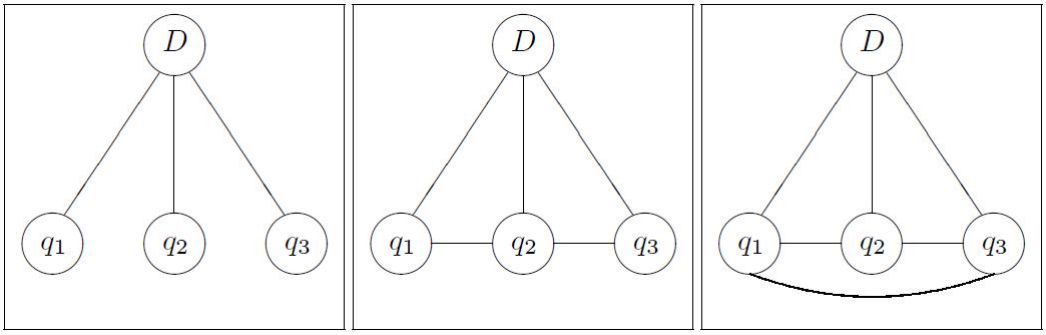

Figure 5.2: Markov random fields: Full Independence (left); Sequential Dependence (center); Full Dependence (right) from [Metzler and Croft, 2005]

Formally, an undirected graph G containing nodes that represent random variables, and the edges define the independence semantics between the random variables, i.e., D and Q in this case. To rank the documents for a given query, a ranking expression based on the posterior can be found:

$$\begin{aligned} P_\Lambda(D|Q) &= \tfrac{P_\Lambda(Q,D)}{P_\Lambda(Q)} \\ P_\Lambda(D|Q) &\propto \log P_\Lambda(Q,D) - \log P_\Lambda(Q) \\ &\propto \textstyle\sum_{c \in C(G)} \log \psi(c;\Lambda), \end{aligned}$$

which can be computed efficiently for reasonable graphs, and where the potential functions $\psi(c;\Lambda)$ are commonly parameterized as:

$$\psi(c;\Lambda) = exp[\lambda_c f(c)],$$

where $f(c)$ is some real-valued *feature function* over clique values and λ_c is the weight given to that particular feature function. The ranking function then becomes:

$$P_\Lambda(D|Q) \propto \sum_{c \in C(G)} \lambda_c f(c). \tag{5.16}$$

The ranking expression given in equation (5.16) provides great flexibility in choosing the feature functions. The feature functions that are chosen for use depend on the application and choice of the designer. This means the model can use almost any type of information available (document term frequencies, term-term co-occurrences, meta data, etc).

The three feature functions deemed of greatest interest by Metzler and Croft are the full independence, sequential dependence and full dependence variants, shown in Figure 5.2.

The *full independence* (FI) variant assumes independence between query terms. The potential function used to evaluate this variant, when using a smoothed language modeling approach, is:

$$\begin{aligned}\psi_T(c) &= \lambda_T \log P(q_i|D) \\ &= \lambda_T \log \left[(1-\alpha_D) \frac{tf_{q_i,D}}{|D|} + \alpha_D \frac{cf_{q_i}}{|C|} \right], \end{aligned} \tag{5.17}$$

where α_D as the smoothing parameter, $tf_{q_i,D}$ is the number of times term q_i occurs in document D, $|D|$ is the length of document D, cf_{q_i} is the frequency of term q_i in the entire collection, and $|C|$ is the term count for the collection. It can be seen that the FI variant is effectively the unigram language model with JM smoothing.

The *sequential dependence* (SD) variant assumes dependence between terms in the query. In this variant, the document dependent cliques form contiguous sequences of query terms. For example, if a query is *train station security measures*, the model will also include estimates for n-grams formed the query, including *train station*, *train station security*, *train station security measures*, *station security*, *station security measures* or *security measures*. The presence of any of these cliques in a document suggests strong evidence of relevance. The potential function for the SD variant, when using a JM smoothed language modeling approach, has the following form:

$$\begin{aligned}\psi_O(c) &= \lambda_O \log P(\#1(q_i, \ldots, q_{i+j})|D) \\ &= \lambda_O \log \left[(1-\alpha_D) \frac{tf_{\#1(q_i,\ldots,q_{i+j}),D}}{|D|} + \alpha_D \frac{cf_{\#1(q_i,\ldots,q_{i+j})}}{|C|} \right], \end{aligned} \tag{5.18}$$

where $tf_{\#1(q_i,\ldots,q_{i+j}),D}$ is the number of times the phrase $(q_i, \ldots, q_{i+j})$ of length j appears in document D, and $cf_{\#1(q_i,\ldots,q_{i+j})}$ is the in-collection frequency of the same phrase. From Equation (5.18), it can be seen that for a query of length n the number of phrases (cliques) to be computed for the feature function of the SD variant is $\sum_{i=1}^{n-1} i$ or $S_n = n - 1$. This would have computational costs that may restrict the application of the variant to short queries. Another approach to addressing the polynomial complexity of the SD variant is proposed by Metzler and Croft [2005] and relates to the use of an external data source (e.g., Wikipedia) to identify the most effective n-grams, ignoring all other possible n-grams contained within the query.

The third variation, the *full dependence* (FD) model, assumes complete dependence over cliques that consist of two or more query terms in any order $q_i, \ldots, q_l$ and the document node D. This means any combination of query terms will create a clique to be evaluated. The

potential function for the FD variant has the following form:

$$\begin{aligned}\psi_U(c) &= \lambda_U \log P(\#uwN(q_i, \ldots, q_l)|D) \\ &= \lambda_U \log \left[(1-\alpha_D) \frac{tf_{\#uwN(q_i,\ldots,q_l),D}}{|D|} + \alpha_D \frac{cf_{\#uwN(q_i,\ldots,q_l)}}{|C|} \right], \end{aligned} \tag{5.19}$$

where $tf_{\#1(q_i,\ldots,q_l),D}$ and $cf_{\#1(q_i,\ldots,q_l),D}$ are the number of times the terms $\{q_i, \ldots, q_l\}$ appear ordered or unordered in a window of N terms within the document and collection respectively. The FD variant has even greater computational complexity than the SD variant.

In practice, query languages are introduced to more easily limit the dependencies to meaningful parts of the query, e.g., manual selection of useful n-grams in the SD variant can be specified in a query language to overcome the computational complexity of estimating all possible query n-grams. The MRF model has been shown to provide significant improvements in retrieval effectiveness when compared to a number of models that ignore term dependencies, including the unigram language model.

The use of a Markov random field within the MRF model to formally combine features was the inspiration behind its use within the TE model. However, using a Markov random field is unrelated to the types of information being modelled by the features. This can be seen by considering the type of word associations underpinning the estimation techniques, i.e., Equation (5.17) models no dependencies; Equation (5.18) and Equation (5.19) model syntagmatic associations; and the TE model in Equation (3.32) models both syntagmatic and paradigmatic associations. This highlights the fact that the Markov random field provides a convenient mathematical tool to formalise the combination of features.

5.2.3 Summary

The overview of information retrieval models presented thus far, have highlighted the heavy statistical nature of existing approaches. This includes the modelling of term dependencies within the MRF model. Even though the SD and FD variants of the MRF model, along with other dependency based models [Tao and Zhai, 2007] could be argued to use information about syntagmatic associations, they do not appear to be motivated from any specific linguistic theory. Previous attempts to use linguistically motivated approaches to modelling term dependencies have focused on using semantic information about query terms to augment the query representation within the information retrieval process [Bai et al., 2005, Voorhees, 1994].

As shown in Figure 5.1, the information retrieval process judges the relevance of a document for a given query by comparing the representation of the document to that of the query. The query is commonly accepted to be an imprecise representation of the user's information need. The cognitive process undertaken by a user in formulating their query from their information need, or in reformulating the query after reviewing the search results, is likely to be heavily dependent on the meaning of words. Therefore, it can be argued that a theoretically grounded application of the TE model of word meaning would be to expand the query representation within the information retrieval process. To understand how this could be achieved, a review of a number of relevant query expansion techniques will be presented.

5.3 Query Expansion

Ever since the Cranfield experiments in document retrieval during the 1960's it has been well known that a query is an imprecise description of the user's real information need [Cleverdon et al., 1966]. For this reason there has been, and still is, a strong interest in the use of *query expansion* techniques [Lv and Zhai, 2010, Metzler and Croft, 2007, Xu and Croft, 1996]. These techniques allow the original query to be augmented to create a more precise representation of the information need and have been shown to increase average retrieval effectiveness [Lv and Zhai, 2010, Metzler and Croft, 2007].

Figure 5.1, illustrates how information about the retrieved documents can be used to augment the query representation. This query expansion process is often achieved using relevance feedback, which relies on the user indicating which of the top k returned documents were relevant. To reduce the burden on the user the top k documents can be assumed to be relevant, and in this case, the relevance feedback setting is referred to as pseudo relevance feedback or blind feedback.

Query expansion within a (pseudo) relevance feedback setting has been shown to provide significant improvements in retrieval effectiveness [Lavrenko, 2004, Lv and Zhai, 2010, Metzler and Croft, 2007]. However, this process is often sensitive to model parameter tuning, and does not consistently assist retrieval effectiveness for all queries [Billerbeck and Zobel, 2004, Collins-Thompson, 2009].

Query expansion techniques are also specific to the mathematical framework of the underlying document retrieval model. Even though our review of document retrieval models have

focused on probabilistic models, as these dominate those classed as state-of-the-art, we will provide an example of a popular and successful vector based query expansion technique as it will be referred to in Section 8.4 when looking at future applications of the TE model.

5.3.1 Rocchio

The Rocchio method [Rocchio, 1971] is one of the most popular and successful query expansion techniques designed for working with geometric representations, such as those found within vector space models [Berry et al., 1999, Buckley, 1995, Salton and Buckley, 1990, Salton et al., 1975]. Rocchio updates the query vector weights using the relevance information, such that the query vector is moved closer in space to the vectors representing the relevant documents and away from those representing non-relevant documents. The most common form of the Rocchio algorithm modifies the initial query weights of the query vector Q, according to:

$$q_j(1) = \alpha q_j(0) + \beta \frac{1}{|R|} \sum_{D_i \in R} d_{ij} - \gamma \frac{1}{|NR|} \sum_{D_i \in NR} d_{ij}, \tag{5.20}$$

where $q_j(0)$ is the initial weight of term j, R is the set of relevant documents in the collection, d_{ij} is the weight of term j in document D_i, NR is the set of non-relevant documents in the collection, and α, β and γ are parameters that control the effect of each component in the equation. In particular, β influences the amount of positive feedback used and γ influences the amount of negative feedback used.

5.3.2 The Relevance Modelling Framework

Rocchio cannot directly augment query representations within probabilistic document retrieval models, as they are probability distributions. However, the ideas behind the Rocchio approach have been used to create a model-based feedback technique that minimises the divergence between the query distrbution and those of the (pseudo) relevant documents [Zhai and Lafferty, 2001].

Another popular technique, that formally augments query representations within the language modelling framework, is known as the relevance modelling approach [Lavrenko and Croft, 2001]. This approach is robust [Lv and Zhai, 2009a] and is regarded as a benchmark in query expansion research [Lv and Zhai, 2010, Metzler and Croft, 2007] and hence will be used as the reference approach for comparison with techniques developed in this thesis.

The relevance model augments the query representations used within the information retrieval process by estimating the probability of observing a word w given some relevant evidence for a particular information need, represented by the query Q:

$$\begin{aligned} P(w|Q) = P(w|R) &= \int_D P(w|D)P(D|Q), \\ &\approx \frac{\sum_{D \in \mathcal{R}_Q} P(w|D)P(Q|D)P(D)}{\sum_w \sum_{D \in \mathcal{R}_\mathcal{Q}} P(w|D)P(Q|D)P(D)}, \end{aligned} \tag{5.21}$$

where $\mathcal{R}_\mathcal{Q}$ is the set of documents pseudo-relevant or relevant to query Q, D is a document in $\mathcal{R}_\mathcal{Q}$, $P(D|Q)$ is the document score of D given Q produced by the underlying language model, and $P(w|D)$ is estimated based on document statistics. To simplify the estimation, $P(D)$ is assumed uniform over this set of documents and thus can be ignored for rank equivalent reasons. When sampled across the vocabulary of terms within the model, an updated multinomial query model is produced, and used to re-rank the documents.

The Unigram Relevance Model

In the unigram variant of the relevance model, where term dependencies are not explicitly modelled, $P(w|D)$ is often estimated based on the Dirichlet smoothed query likelihoods, shown in Equation (5.15).

The relevance model estimate $P(w|R)$ is often interpolated with the original query model estimate, to form a final estimate:

$$P(w|Q) = \alpha P_o(w|Q) + (1 - \alpha)P(w|R), \tag{5.22}$$

where α is the feedback interpolation coefficient that determines the mix with the original query model estimate $P_o(w|Q)$. This form of unigram based relevance model, in combination with Equation (5.15) is referred to as **RM3** in this research.

Even though the unigram relevance model has demonstrated significant improvements in retrieval effectiveness over a unigram language model, recent research has demonstrated that significant improvements can be made over the unigram relevance model by incorporating explicit information about term dependencies into the expansion process [Lv and Zhai, 2010, Metzler and Croft, 2007]. These approaches include the *positional relevance model* (PRM) [Lv and Zhai, 2010] and *latent concept expansion* (LCE) [Metzler and Croft, 2007].

The Positional Relevance Model

Lv and Zhai [2010] found that using positional information of terms within the relevance modelling framework can significantly improve retrieval effectiveness over a unigram approach. Based on the intuition that topically related content is grouped together in text documents, the positional relevance model (PRM) uses proximity and positional information to form expansion term estimates to update the query model. The estimates for expansion terms are computed as:

$$P(w|Q) = \frac{P(w,Q)}{P(Q)} \propto P(w,Q) = \sum_{D \in \mathcal{R}_Q} \sum_{i=1}^{|D|} P(w,Q,D,i), \tag{5.23}$$

where i indicates a position in document D, and $\mathcal{R}_Q$ is the set of feedback documents (assumed to be relevant). Two sampling methods are proposed to estimate $P(w,Q,D,i)$. The first uses independent and identically distributed (*iid*) sampling, such that:

$$P(w,Q,D,i) \propto \frac{P(Q|D,i)P(w|D,i)}{|D|}. \tag{5.24}$$

The second approach to estimating $P(w,Q,D,i)$ uses conditional sampling, such that:

$$P(w,Q,D,i) = P(Q)P(D|Q)P(i|Q,D)P(w|D,i). \tag{5.25}$$

Both approaches are based on the following estimate of $P(w|D,i)$:

$$P(w|D,i) = (1-\lambda)\frac{c'(w,i)}{\sqrt{2\pi\sigma^2}} + \lambda P(w|C) \tag{5.26}$$

where

$$c'(w,i) = \sum_{j=1}^{|D|} c(w,j) \exp\left[\frac{-(i-j)^2}{2\sigma^2}\right],$$

and $c(w,j)$ is the *actual* count of term w at position j, $|D|$ is the length of the document, λ is a smoothing parameter and σ is used to parameterize the Gaussian kernel function ($\frac{-(i-j)^2}{2\sigma^2}$).

Both variants of the positional relevance model were reported to achieve significantly superior retrieval effectiveness over the unigram relevance model, with the *iid sampling* approach performing slightly better than the conditional sampling approach [Lv and Zhai, 2010].

5.3.3 Latent Concept Expansion

Latent concept expansion (LCE) [Metzler and Croft, 2007] was developed as a query expansion approach within the MRF document ranking model. LCE formally combines various likelihood based features to produce an estimate for expansion concepts (i.e., potentially n-tuples) through the use of a Markov random field.

For the simplest variant of LCE, in which concepts are made up of single terms, the expansion term likelihoods can be computed as:

$$
\begin{aligned}
&P_{H,\Lambda}(e|Q) \propto \\
&\sum_{D \in \mathcal{R}_\mathcal{Q}} \exp[\lambda_{T_D} \sum_{w \in Q} \log \left[(1-\alpha) \frac{tf_{w,D}}{|D|} + \alpha \frac{cf_w}{|C|} \right] + \\
&\qquad \lambda_{O_D} \sum_{B \in Q} \log \left[(1-\beta) \frac{tf_{\#1(B),D}}{|D|} + \alpha \frac{cf_{\#1(B)}}{|C|} \right] + \\
&\qquad \lambda_{U_D} \sum_{B \in Q} \log \left[(1-\beta) \frac{tf_{\#uw(B),D}}{|D|} + \beta \frac{cf_{\#uw(B)}}{|C|} \right] + \\
&\qquad \log \frac{\left((1-\alpha) \frac{tf_{e,D}}{|D|} + \alpha \frac{cf_e}{|C|} \right)^{\lambda'_{T_D}}}{\left(\frac{cf_e}{|C|} \right)^{\lambda'_{T_Q}}}], \qquad (5.27)
\end{aligned}
$$

where $B \in Q$ denotes the set of bigrams formed from query Q, cf_w is the collection frequency of term w, $tf_{w,D}$ is the term frequency of term w in document D, α and β are smoothing parameters, $\lambda_{T_D}, \lambda_{O_D}, \lambda_{U_D}, \lambda'_{T_D}, \lambda'_{T_Q}$ are model hyper-parameters, and $\#1(b)$ and $\#uw(b)$ are Indri operators [Ogilvie and Callan, 2001] that represent ordered bigram and unordered bigram counts, respectively. Bigram information has been used successfully in the past to estimate query expansion terms [Miller et al., 1999]. It is important to note that the query likelihood estimates in Equation (5.27) predominantly model syntagmatic information, and no consideration of paradigmatic information is given.

When LCE was used to augment query representations (i.e., within the MRF document ranking model) it was shown to provide significantly improved retrieval effectiveness over a unigram relevance model. However, since the MRF document ranking model has been shown to provide significant improvements in retrieval effectiveness over the unigram language model, which underpins the unigram relevance model [Metzler and Croft, 2005] this comparison is likely to be unreliable. This is because the estimation techniques of each are based on two different document ranking models, which will produce two unique sets of (pseudo) relevant documents.

In addition, when evaluating models that employ many free parameters, i.e., increasing the degrees of freedom within the model, it can be difficult to determine the extent to which improvements in task performance are because of the increased degrees of freedom (i.e., retrieval effectiveness when referring to query expansion) [Metzler and Zaragoza, 2009]. The lesson from this discussion indicates that a rigorous evaluation of query expansion techniques requires

that same document ranking model be used to underpin each technique, i.e., ensuring estimates are based on the same (pseudo) relevant documents, and that the number of free parameters between models should be kept the same, or very similar.

Parameterized Concept Weighting

Parameterized concept weighting [Bendersky et al., 2011b] combines LCE with a learning to rank algorithm and other external information sources, to provide further improvements in retrieval effectiveness. The ability to continually combine techniques such as query expansion, learning to rank, spam filtering and external information sources is commonly used to achieve superior performance in real-world, industry-like settings [Bendersky et al., 2011a, Boystov and Belova, 2011]. However, as these approaches do not have independent variables rigorous scientific evaluation becomes problematic.

5.4 Summary

Even though the query expansion approaches presented here are statistically motivated, they can be argued to implicitly or explicitly access information about syntagmatic associations between words. The implicit use occurs due to the (pseudo) relevance feedback process itself, where words that co-occur more often with the query terms are more likely to exist within the set of (pseudo) relevant documents from which the expansion term estimates are derived. The explicit use stems from the modelling of positional information (i.e., in the positional relevance model), or access to the relative frequencies of n-grams within the SD variant of LCE.

Linguistically motivated approaches to query expansion have been presented in the past, but have been met with mixed success [Bai et al., 2005, Bruza and Song, 2002, Grefenstette, 1992, Hoenkamp et al., 2010, Voorhees, 1994]. One attempt that was unable to produce consistent improvements in retrieval effectiveness used primariy paradigmatic information, sourced from an external linguistic resource (i.e., WordNet) [Voorhees, 1994]. WordNet is an ontology that groups together nouns, verbs adjectives and adverbs in sets, known as *synsets*, based their on cognitive synonymy. Synsets are linked by conceptual semantic and lexical relations and model a strict form of paradigmatic information.

Two other attempts using linguistically motivated techniques include the Information Flow [Bruza

and Song, 2002] and epiHAL [Hoenkamp et al., 2010] models, which both model some mix of syntagmatic and paradigmatic information using a HAL-based model (Section 2.5.3). These approaches did demonstrate improvements in retrieval effectiveness on a number of small data sets. However, given these approaches did not include the recent advances in SSM technologies that exist within the TE model, we argue further improvements in retrieval effectiveness may be possible using a query expansion approach based on the TE model.

The use of the TE model within query expansion will differ in many ways from these previous linguistically motivated attempts. Firstly, it will explicitly control the mix of syntagmatic and paradigmatic information being used in the estimation process. This will allow a more rigorous evaluation of the benefits of each type of linguistic association for query expansion to be made.

Secondly, the TE model is based on the definition of meaning proposed within structural linguistics, of which a *cognitive* impression on a user is an intuitive aspect (Section 2.2). This cognitive aspect of the TE model fits well within the growing body of *information seeking and retrieval* (ISR) research that portrays the user as an integral part of the process [Ingwersen, 1992, Ingwersen and Järvelin, 2005].

Thirdly, the development of a new query expansion technique based on the TE model will be formalised within the relevance modelling framework, as opposed to the heuristic expansion techniques used in previous approaches based on SSM technology. The following chapter outlines how this is achieved.

Chapter 6

The Tensor Query Expansion (TQE) Approach

The previous chapter highlighted the statistical nature of current state-of-the-art information retrieval models, including techniques used within query expansion. Query expansion techniques that model term dependencies have demonstrated significant improvements in retrieval effectiveness over approaches that ignore term dependencies. However, these dependency based approaches only use information about half the associations (i.e., syntagmatic) that give rise to word meanings. The user's dependence on word meanings when formulating their query, motivates the use of the TE model to access information about both syntagmatic and paradigmatic associations within the query expansion process.

This chapter provides a formalism for placing the TE model within the relevance modelling framework. This framework was chosen as it formally augments query representations within the language modelling framework, which is a formal and extensively researched document ranking model that demonstrates state-of-the-art performance.

6.1 A Linguistically Motivated Relevance Model

One of the strengths of the formal framework underpinning the TE model, is the ability to extend it. Because of this, the framework presented in Chapter 3 can be updated to support the process involved in augmenting query representations in the relevance modelling framework. This augmentation process can briefly be described as augmenting an initial query representation that consists of a sequence of query terms $Q = (q_1, \ldots, q_p)$ with a set of associated terms. Using the TE model to augment query representations will involve using a mix of syntagmatic and paradigmatic information relating to the query terms to identify these associated terms.

We propose a formalism that uses a second-order TE model, and hence assumes expansion terms are single words, not n-grams or n-tuples ($n > 1$). The choice to restrict the modelling between query terms and expansion terms to single words is supported by past research, which has found that the use of multi-word concepts for expansion does not provide consistent improvements in retrieval effectiveness when compared to methods using single word expansion terms [Carpineto and Romano, 2012, Metzler and Croft, 2007]. However, researchers believe multi-word concepts should help overcome vocabulary issues Bai et al. [2007], and that their inconsistent performance may be overcome in the future [Carpineto and Romano, 2012]. Because of this result, any investigation into the use of higher-order TE models to support multi-word expansion concepts is left for future work.

Extending the original TE model to support query expansion is formalised within a probabilistic graphical model framework [Koller and Friedman, 2009]. Let an undirected graph G contain nodes that represent the random variables (Q, V_k, w), and let the edges define the independence semantics between the random variables, as shown in Figure 6.1. Term w is constrained to exist within the vocabulary V_k. Within the graph, a random variable is independent of its non-neighbours given observed values of its neighbours.

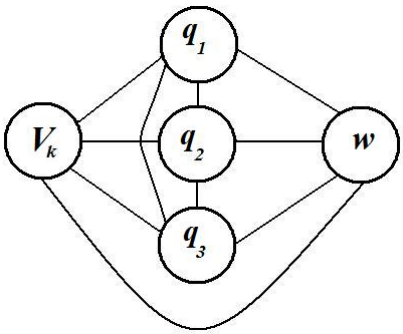

Figure 6.1: Example of a graphical model for a three term query.

We parameterize the graph based on clique sets to provide more flexibility in encoding useful features over cliques in the graph. The joint distribution over the random variables in G is defined by:

$$P_{G,\Gamma}(Q, w, V_k) = \frac{1}{Z_\Gamma} \prod_{c \in cl(G)} \varphi(c; \Gamma), \tag{6.1}$$

where $Q = q_1, \ldots, q_p$, $cl(G)$ is the set of cliques in G, each $\varphi(.; \Gamma)$ is a non-negative *potential function* over clique configurations parameterized by Γ, and $Z_\Gamma = \sum_{Q,w} \prod_{c \in cl(G)} \varphi(c; \Gamma)$

normalizes the distribution. The joint distribution is uniquely defined by the graph G, potential functions φ and the parameter Γ. Because the logarithm of products is equal to the sum of logarithms, the joint distribution can be simplified to obtain:

$$\log P_{G,\Gamma}(Q, w, V_k) = \frac{1}{Z_\Gamma} \sum_{c \in cl(G)} \log \varphi(c; \Gamma). \tag{6.2}$$

The potential functions in Equation (6.2) are commonly parameterized as:

$$\varphi(c; \Gamma) = \exp[\gamma_c f(c)], \tag{6.3}$$

where $f(c)$ is a real-valued *feature function* over clique values and γ_c is the weight given to that particular feature function. Substituting Equation (6.3) into Equation (6.2) gives:

$$\log P_{G,\Gamma}(Q, w, V_k) = \frac{1}{Z_\Gamma} \sum_{c \in cl(G)} \gamma_c f(c). \tag{6.4}$$

After G is constructed, we can compute the conditional probability of an expansion term w given Q, as:

$$P_{G,\Gamma}(w|Q) = \frac{P_{G,\Gamma}(Q, w, V_k)}{\sum_{w \in V_k} P_{G,\Gamma}(Q, w, V_k)}, \tag{6.5}$$

where V_k is the universe of all possible vocabulary terms and w is a possible expansion term. Note that the denominator in Equation (6.5) is constant for all terms in the vocabulary.

By using Equation (6.4) and Equation (6.5) with constant terms removed, a rank equivalent form for the conditional probability can be written as:

$$P_{G,\Gamma}(w|Q) \propto \sum_{c \in cl(G)} \gamma_c f(c), \tag{6.6}$$

where a constraint of $\sum_{c \in cl(G)} \gamma_c = 1$ is applied for ease of training.

6.1.1 Model Parameterization

The conditional probability expressed in Equation (6.6) provides a formal method for combining feature functions that will allow measures of syntagmatic and paradigmatic associations to be modelled via cliques in the graph. For the graph shown in Figure 6.1 a number of useful clique sets capturing these associations are summarised in Table 6.1.

Substituting the syntagmatic and paradigmatic measures as feature functions applied over appropriate cliques, as shown in Table 6.1, the expression in Equation (6.6) can be expanded, giving:

$$P_{G,\Gamma}(w|Q) \propto \gamma_{T_{\text{syn}}} s_{\text{syn}}(Q, w) + \gamma_{T_{\text{par}}} s_{\text{par}}(Q, w), \tag{6.7}$$

Set	Description
T_{par}	Set of cliques containing the vocabulary node and exactly one query term node and the expansion term (w) node. This set does not include edges between the query term node and expansion term node.
T_{syn}	Set of cliques containing the vocabulary node connected to all query nodes by edges, and the vocabulary node to the expansion term (w) node connected by an edge.

Table 6.1: Summary of the TQE clique sets to be used.

where $\gamma_{T_{\text{syn}}}, \gamma_{T_{\text{par}}} \in [0,1]$, $\gamma_{T_{\text{syn}}} + \gamma_{T_{\text{par}}} = 1$, and $s_{\text{syn}}(Q,w)$ is the chosen syntagmatic feature. By normalising the distribution and replacing $\gamma_{T_{\text{syn}}}$ and $\gamma_{T_{\text{par}}}$ with a single interpolation parameter, γ, the rank equivalent estimate in Equation (6.7) can be rewritten as a conditional:

$$P_{G,\Gamma}(w|Q) = \frac{1}{Z_\Gamma}\left[(1-\gamma)s_{\text{syn}}(Q,w) + \gamma s_{\text{par}}(Q,w)\right], \tag{6.8}$$

where $\gamma \in [0,1]$ mixes the amount of syntagmatic and paradigmatic features used in the estimation, and Z_Γ is used to normalise the distribution.

The estimate in Equation (6.8) is produced from a multinomial distribution, akin to those in the unigram and positional relevance models (Section 5.3.2), and hence can be used to augment the query representations within the language modelling framework. Using the relevance models feedback interpolated form, shown in Equation (5.22), the final conditional probability becomes:

$$P(w|Q) = \alpha P_o(w|Q) + (1-\alpha)P_{G,\Gamma}(w|Q). \tag{6.9}$$

The construction of this linguistically motivated relevance model ensures that modifying the mixing parameter (γ) will add paradigmatic information to the query expansion process in a controlled manner. Assuming that all other parameters in the system are controlled for, the influence of syntagmatic and paradigmatic information on retrieval effectiveness can be robustly evaluated within this framework. This linguistically motivated relevance model is referred to as *tensor query expansion* (TQE) in the remainder of this work.

6.2 Choosing Syntagmatic and Paradigmatic measures

The TE framework provides great flexibility when choosing the similarity measures to use when modelling syntagmatic and paradigmatic associations within the model. The following section describes the selection process undertaken when choosing two measures for modelling syntagmatic and paradigmatic associations within the TQE approach. The measures will need to model associations between a sequence of query terms, represented as $Q = (q_1, \ldots, q_p)$ and a possible expansion term w, identified in Equation (6.9).

6.2.1 Modelling Syntagmatic Associations

The original syntagmatic measure in Equation (3.27) was used to model syntagmatic associations between two vocabulary terms q and w for the semantic tasks outlined in Chapter 4. However, for query expansion where a sequence of query terms exists, the more general form of the syntagmatic measure, shown in Equation (3.26) can be used and is restated here:

$$s_{\text{syn}}(Q, w) = \frac{\sum_{j \in \{V|w \in Q\}} f_{jw}^2 + \sum_{j \in \{V|w \in Q, j \neq w\}} f_{wj}^2 + \sum_{i \in \{Q|i \neq w\}} (f_{iw}^2 + f_{wi}^2)}{\sqrt{\sum_{i \in Q} \left[\sum_{j \in V} f_{ji}^2 + \sum_{j \in \{V|j \neq i\}} f_{ij}^2\right]} \sqrt{\sum_{j \in V} f_{jw}^2 + \sum_{j \in \{V|j \neq w\}} f_{wj}^2}}, \quad (6.10)$$

where $Q = q_1, \ldots, q_p$ are the query terms in Q, f_{ab} is the co-occurrence frequency of term a appearing before term b in the vocabulary, and V_k is the set of vocabulary terms. It is worth noting that this measure of syntagmatic association relates to the second-order TE model, which captures word order and co-occurrence information of single words in memory matrices (refer to Section 3.1.1). As mentioned in Section 6.1, for the application of query expansion it will be assumed that the syntagmatic and paradigmatic associations formed between single words will be sufficient to achieve significant improvements in retrieval effectiveness.

Past query expansion research has also found that effective syntagmatic associations exist between words far apart in natural language [Xu and Croft, 1996], and suggests that a much wider context window should be used in modelling syntagmatic associations within the $s_{\text{syn}}(,)$ feature of the TQE approach. This result was tested by evaluating the retrieval effectiveness of the $s_{\text{syn}}(,)$ feature in Equation (6.10) for various context window radii (Figure 6.2), using the experimental setup outlined in Section 7.1).

Figure 6.2 confirm that substantial improvements in *mean average precision* (MAP, refer to Section 5.2.1) can be achieved on the G2 and CW data sets (Table 7.1) when a wider context

window is used to model syntagmatic associations. This result was also obtained for the AP, WSJ and ROB data sets in Table 7.1. Figure 6.2 also illustrates the robustness of $s_{\text{syn}}(,)$ for context window radii above 200.

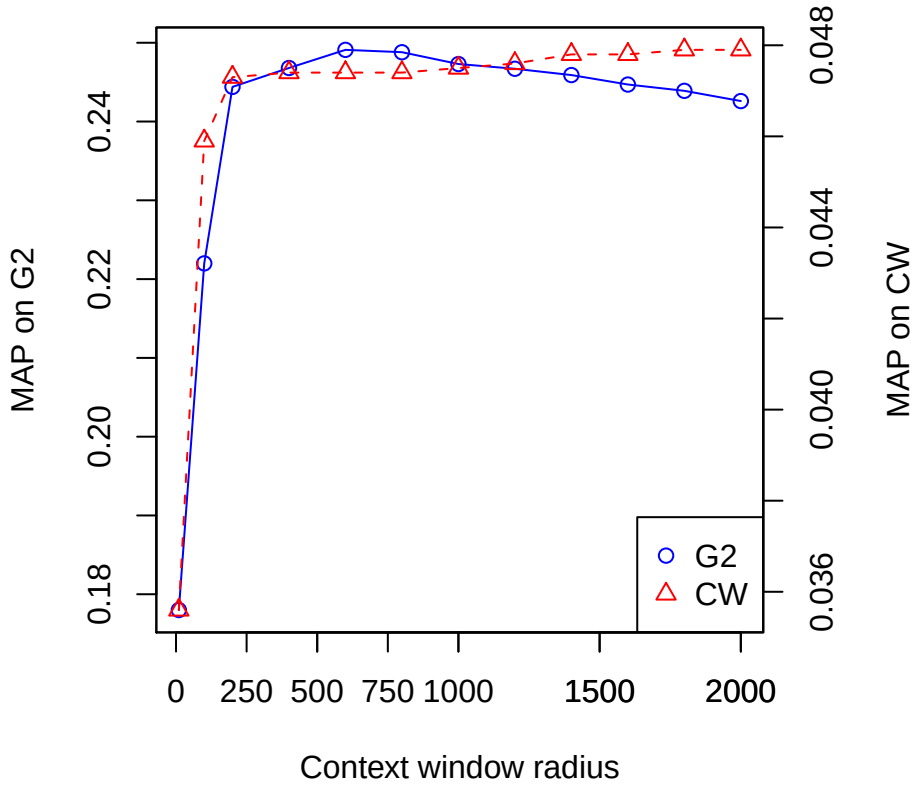

Figure 6.2: Sensitivity of the $s_{\text{syn}}(,)$ measure with respect to context window radius, on the G2 and CW data sets.

As the context window in the TE binding process does not cross document boundaries, it is worth considering the retrieval effectiveness achieved when the context window radius is set to each document's length. Using this radius, the MAP scores achieved by $s_{\text{syn}}(,)$ in Equation (6.10) are 0.2491 and 0.0492 on the G2 and CW data sets, respectively. This result indicates that superior retrieval effectiveness is achieved using the whole document as the context window to compute syntagmatic associations.

To achieve this the context window radius (r) should be set to the document length ($|D|$) in the TE binding process. Making this substitution, the binding process of the second-order TE model defined in Equation (3.6) becomes:

$$\boldsymbol{M}_w = \sum_{t \in CW}^{t \prec w} (|D| - d_t).\boldsymbol{e}_t \otimes \boldsymbol{e}_w^T + \sum_{t \in CW}^{t \succ w} (|D| - d_t).\boldsymbol{e}_w \otimes \boldsymbol{e}_t^T. \tag{6.11}$$

The algebraic form of elements on row w of the matrix M_w in Equation (6.11) becomes:

$$f_{wj} = \sum_{D \in \{\mathcal{R}_\mathcal{Q} | w \in D\}} df_j(|D| - \bar{d}_{w,j}), \tag{6.12}$$

and on column w:

$$f_{iw} = \sum_{D \in \{\mathcal{R}_\mathcal{Q} | w \in D\}} df_i(|D| - \bar{d}_{i,w}), \tag{6.13}$$

where D is the set of terms making up a document in the set of pseudo relevant documents $\mathcal{R}_\mathcal{Q}$, $|D|$ is the length of document D, df_j is the frequency of term j in document D, $\bar{d}_{w,j}$ is the average number of terms separating term w from term j when w is seen before j in document D, and $\bar{d}_{i,w}$ is the average number of terms separating term w from term i when w is seen after i in document D.

When Equations (6.12) and (6.13) are substituted into Equation (6.10), the syntagmatic feature $s_{\text{syn}}(,)$ produces higher scores for terms that occur frequently (large df_j and df_i) in the pseudo relevant documents. This is a similar result to that produced by the Dirichlet smoothed likelihood estimation in Equation (5.21) that underpins RM3. However, Equation (5.21) contains a document normalisation factor. The cosine metric that defines $s_{\text{syn}}(,)$, also uses a form of normalisation that is linked to the document length. Equations (6.12) and (6.13) infer that terms that occur in larger documents will likely produce larger Frobenius norms (i.e., the denominator of Equation (6.10) becomes larger for larger documents), and hence the syntagmatic measure is effectively normalised by the document length.

Therefore, the estimation techniques used in RM3 and $s_{\text{syn}}(,)$, when the binding process in Equation (6.11) is used, are effectively based on term document frequencies and a document length normalisation factor. This result would lead us to believe that RM3 may be using very similar information to TE model's syntagmatic feature. This is an intuitive result since the unigram relevance model effectively updates the query model to *enhance* terms that appear often in the set of k (pseudo) relevant documents. In a unigram language model the top k retrieved documents that make up the set of (pseudo) relevant documents are those that contain the query terms most often. Therefore, these enhanced terms are those that co-occur often with the query terms, and hence have strong syntagmatic associations effectively modelled with a context window radius equal to the document length.

To confirm this intuition the overlap between the sets of expansion terms used by the best performing unigram relevance model (RM3) and TQE (using $\gamma = 0$) was undertaken on the G2 and CW data sets outline in Table 7.1 and have been reported by Symonds et al. [2012a].

This investigation found that on average the best performing RM3 and TQE using $\gamma = 0$ and the syntagmatic feature $s_{\text{syn}}(.)$ defined in Equation (6.10) had on average 19 out of a possible 30 expansion terms in common. This suggests that RM3 is using quite a lot of syntagmatic information when estimating expansion terms.

Given this finding, basing a measure of syntagmatic associations on the estimation technique used within the unigram relevance model has many advantages. From a computational complexity perspective, the advantages include the fact that no semantic space needs to be built to underpin the syntagmatic measure. The estimation of syntagmatic associations can be made from frequency information already available in the document index.

From an empirical stand point, the advantage of having the TE model use the same method of modelling syntagmatic associations as the unigram relevance model comes from the improvements in experimental variable control. This means that when the mix of syntagmatic and paradigmatic information is modified in TQE, via the gamma (γ) parameter in Equation (6.8), any differences in retrieval effectiveness are due to the influence of the paradigmatic information. For these reasons the unigram relevance model (RM3) was chosen as the benchmark model, and the TE model's syntagmatic measure was based on the same information that underpins the unigram relevance model's estimate.

The resulting measure of the strength of syntagmatic associations between a sequence of query terms, Q and a vocabulary term w used by the TE model in this research is defined as:

$$\begin{aligned} s_{\text{syn}}(Q, w) &= \frac{1}{Z_{\text{syn}}} \sum_{D_i \in V_k(Q)} P(D_i|Q) P(Q|w) \\ &= \frac{1}{Z_{\text{syn}}} \sum_{D_i \in V_k(Q)} s(D_i, Q) \frac{df_w}{|D_i|}, \end{aligned} \tag{6.14}$$

where $V_k(Q)$ is the set of vocabulary terms drawn from the collection of (pseudo) relevant feedback documents, $s(D_i, Q)$ is the document relevance score of the (pseudo) relevant document D_i given query Q, df_w is the document frequency of term w, $|D_i|$ is the length of document D_i and Z_{syn} normalises the distribution. The smoothing parameter (μ) seen in the Dirichlet estimate of Equation (5.15) was ommitted (or effectively set to zero), so that the number of free parameters (degrees of freedom) used in TQE model is the same as the unigram relevance model (RM3), and hence would not be the cause of any differences in retrieval effectiveness between the two approaches.

6.2.2 Modelling Paradigmatic Associations

The original paradigmatic measure, defined in Equation (3.30), was used successfully to model paradigmatic associations between vocabulary terms q and w within the TE model on a number of semantic tasks carried out in Chapter 4. It was later enhanced to include a discounting expression that penalised terms that had strong syntagmatic associations with term q, or concept c_1 when referring to medical concepts, as defined in Equation (4.2). However, to use this enhanced paradigmatic measure for query expansion it needs to support the modelling of paradigmatic associations between a sequence of terms $Q = (q_1, \ldots, q_p)$ and a vocabulary term w. Therefore, Equation (4.2) can be modified to sum the accumulated strengths of term w with each query term, as follows:

$$s_{\text{par}}(Q, w) = \frac{1}{Z_{\text{par}}} \sum_{j \in Q} \sum_{i \in V} \frac{f_{\overline{ij}} . f_{\overline{iw}}}{\max(f_{\overline{ij}}, f_{\overline{iw}}, f_{\overline{wj}})^2}, \tag{6.15}$$

where $f_{\overline{ij}} = (f_{ji} + f_{ij})$ is the unordered co-occurrence frequency of terms i and j, V is the set of vocabulary terms, $\max()$ returns the maximum argument value, and Z_{par} normalizes the distribution of scores, such that $\sum_{w \in V} s_{\text{par}}(Q, w) = 1$. In the context of (pseudo) relevance feedback, the vocabulary V is formed from the set of top k (pseudo) relevant documents.

Effective modelling of paradigmatic associations is often achieved when using a a very small sliding context window during the TE model's binding process in Equation (3.6). This has been highlighted on tasks that rely heavily on paradigmatic information, such as synonym judgement, as seen in Section 4.2. To confirm this finding can be generalised to modelling paradigmatic information within the query expansion process an evaluation of retrieval effectiveness is undertaken for varying context window radii.

This is achieved by using $s_{\text{par}}(,)$ to estimate expansion terms within a pseudo relevance feedback setting using the relevance model framework on the WSJ, G2 and CW data sets outlined in Table 7.1. The retrieval effectiveness achieved by $s_{\text{par}}(,)$ for various context window radii was measured using *mean average precision* (MAP) on each data set, and is shown in Figure 6.3. This graph demonstrates that on average $s_{\text{par}}(,)$ is most effective when a context window radius of 1 is used.

The result of setting $r = 1$ during the TE model's second-order binding process, Equation (3.6), is that the element values in the memory matrices are effectively not scaled and represent the sum of the actual unordered co-occurrence frequencies for the target term within

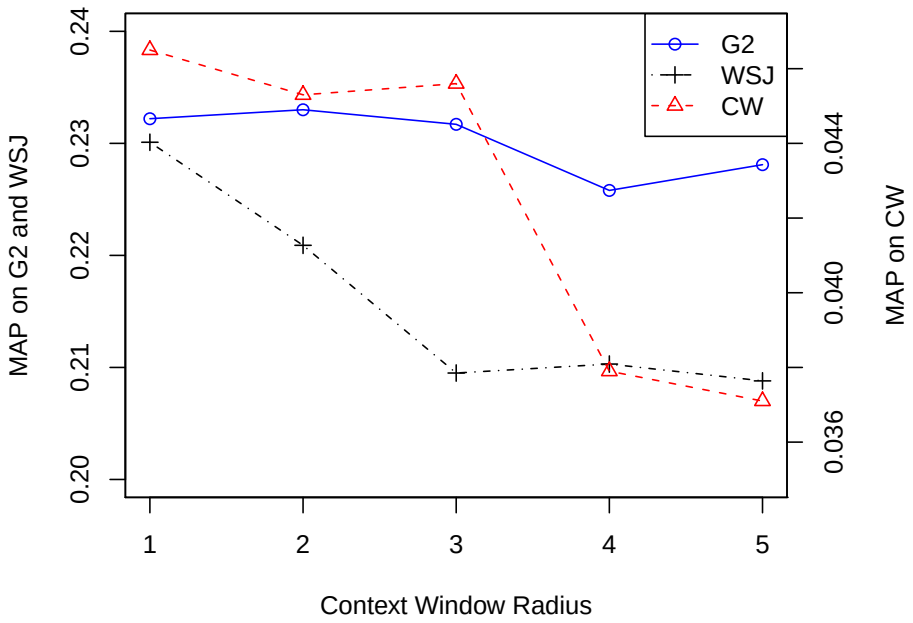

Figure 6.3: Sensitivity of the $s_{\text{par}}(,)$ measure with respect to context window radius, on the WSJ, G2 and CW data sets.

the set of (pseudo) relevant documents. This means that document indexes that contain bigram statistics would be sufficient to model paradigmatic associations, and no additional computations would be required to build the semantic space. However, bigram indexes of the entire document collection require large amounts of storage space, and hence were not used in this research. Therefore, the storage vectors for all vocabulary terms in the set of pseudo relevant documents were constructed. The computational complexity analysis in Section 6.3 assumes storage vectors are used to underpin the paradigmatic measure.

Equation (6.14) and Equation (6.15) define the two measures that will be used to explicitly model syntagmatic and paradigmatic associations, respectively, within the TQE approach developed in this work. The following section provides a computational complexity analysis to allow a comparison of efficiency to be evaluated when comparing TQE with other query expansion approaches.

6.3 Computational Complexity of TQE

The efficiency of the TQE approach can be evaluated by calculating the time and storage costs associated with building the storage vectors and computing the syntagmatic and paradigmatic measures.

The syntagmatic measure, Equation (6.14), is based on the estimation technique used within the unigram relevance model (RM3) and hence will have similar computational complexity; probably less as it does not access global collection statistics (i.e., cf_w), unlike Equation (5.15) used by RM3. Given Equation (6.14) relies on statistics already stored in the document indexes (i.e., df_w and D_i), or calculations from the underlying document model, (i.e., $s(D_i, Q)$), no additional time or storage costs are associated with building the representations. However, the time to compute the syntagmatic measure for each term in the vocabulary, formed from the set of k pseudo relevant documents, is: $T(n) = O(k|V|)$), where $|V|$ represents the size of the vocabulary.

When implementing the paradigmatic measure within the TQE approach, a TE vocabulary needs to be created from the set of pseudo relevant documents, assuming no bigram statistics exist in the document indexes. The process of building a TE vocabulary for the paradigmatic measure involves moving a context window of radius 1 across the text in these documents, and accumulating the co-occurrence frequencies within the storage vectors. As outlined in Section 3.1.4, the time complexity of this TE model's vocabulary building process is determined by considering the worst case, which occurs when the storage vectors are full and a replacement operation is performed. In this case, the basic operation of the binding process becomes a full search of the (T CF) list, giving: $T(n) = O(\frac{D_{SV}}{2})$, where D_{SV} is the storage vector dimensionality.

The time complexity of computing the paradigmatic measure in Equation (6.15) is $T(n) = O(\frac{D^2_{SV_{\text{par}}}}{4}.|Q|)$, where $D_{SV_{\text{par}}}$ is the dimensionality of the storage vector, and $|Q|$ is the length of the query. Therefore, keeping the dimensionality of the storage vector small is important. To test what storage vector dimensionality would provide effective retrieval for the $s_{\text{par}}(,)$ measure, an experiment considering retrieval effectiveness, based on *mean average precision* (MAP), for various storage vector dimensions on the G2, WSJ and CW data sets, outlined in Table 7.1, was undertaken.

Figure 6.4 demonstrates that storage vectors with fewer dimensions (as low as 20) can provide similar if not better retrieval effectiveness than higher dimensions, especially on larger document collections (i.e., G2 and CW). This may be due to the ability of the TMC process to remove low information terms from the storage vectors. The results of this test suggest that storage vectors with lower dimensionality can be used to effectively model paradigmatic associations that assist retrieval effectiveness. This has a considerable efficiency saving when

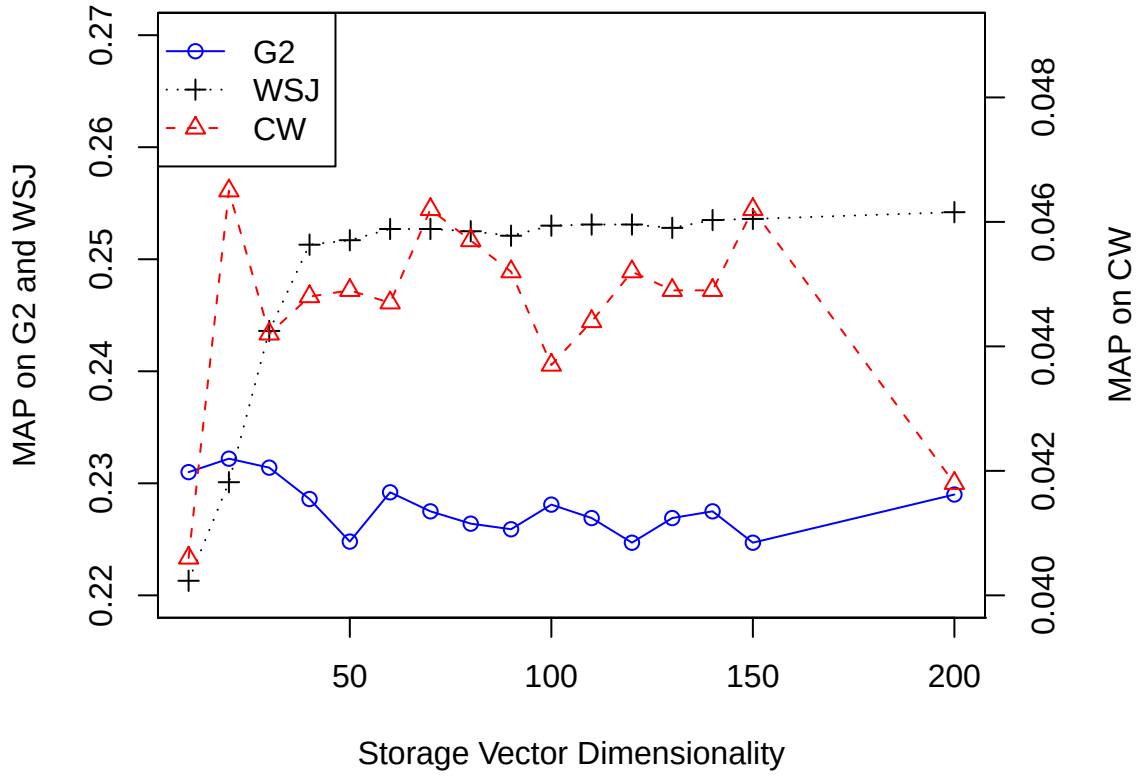

Figure 6.4: Sensitivity of the $s_{\text{par}}(,)$ measure with respect to storage vector dimensionality, on the G2, WSJ and CW data sets, evaluated using MAP to measure retrieval effectiveness.

considering the time complexity of the paradigmatic measure, recall $T(n) = O(\frac{D^2_{SV_{\text{par}}}}{4}.|Q|)$.

Obviously, treating the storage vector size as a free model parameter within TQE and tuning it for each data set would optimise the retrieval effectiveness of TQE. However, to allow a fair comparison between models when evaluating the TQE's performance against relevant benchmarks, the number of free parameters should be kept to a minimum. For the rigorous evaluation of TQE, the aim is to only allow the mix of syntagmatic and paradigmatic information (γ in Equation (6.8)) to be optimised, so that the impact of each on retrieval effectiveness can be more confidently reported. More discussion on this choice is provided in Section 7.1.1. This decision means that choosing a storage vector size that balances retrieval effectiveness with efficiency needs to be made.

Given the importance of larger document collections in modern web retrieval and the efficiency gains from using a storage vector with lower dimension, a storage vector of 20 dimensions was chosen to underpin the TE model's paradigmatic measure. Therefore, the worst case time complexity of the paradigmatic measure in Equation (6.15) is $T(n) = O(\frac{D^2_{SV_{\text{par}}}}{4}.|Q|)$, where $D_{SV_{\text{par}}}$ is the dimensionality of the storage vector, and $|Q|$ is the length of the query.

Therefore, choosing to set $D_{SV_{\text{par}}} = 20$ is likely to provide the best trade off between retrieval effectiveness and efficiency, resulting in a time complexity for $s_{\text{par}}(,)$ of $T(n) = O(100|Q|)$. To update the query representations, an estimate for all terms in the TE model's vocabulary is required, therefore the time complexity of this process would be $T(n) = O(100\overline{|Q|}|V|)$, where $\overline{|Q|}$ is the average query length, and $|V|$ is the size of the TE model's vocabulary.

The storage complexity of the paradigmatic measure is determined by the amount of memory required to hold the storage vectors. That is the number of terms in the vocabulary times the size of the storage vectors, resulting in a storage complexity of $M(n) = O(D_{SV_{\text{par}}}.|V|) = O(20|V|)$. As the syntagmatic measure uses information found within existing document indexes, the storage complexity of the TQE approach is equal to the storage complexity of the paradigmatic measure, i.e., $M(n) = O(D_{SV_{\text{par}}}.|V|) = O(20|V|)$. However, a point to note is that since a context window radius of 1 is used to build the TE vocabulary underpinning the paradigmatic measure, if the document index contained bigram statistics then this could be used instead of buildng the TE vocabulary, and hence no storage complexity difference would exist between RM3 and TQE.

In most cases, no bigram statistics will exist. In these cases, an idea of the actual storage needed for the TQE approach can be gauged by assuming the average vocabulary size created from the (pseudo) relevant documents is 15,000. This is reasonable based on the experiments performed in Chapter 7. Using this value the TQE storage complexity becomes $20 \times 15000 = 30,000$ integers. Assuming a 4 byte integer, the actual memory requirements of the TQE approach amounts to 75kBytes. This is a lot less than would be required by an index containing bigram statistics of the entire collection. Therefore, the use of the TE model's SSM is probably preferred, and the only additional cost to the small amount of storage space would be the time to build the vocabulary, which was shown to be $T(n) = O(\frac{D_{SV}}{2})$.

In total, the time complexity of TQE approach is made up of this time to build the semantic space underpinning the paradigmatic measure, plus the time to compute the syntagmatic and paradigmatic measures across the entire vocabulary. From the previous analysis, this is: $T_{\text{TQE}}(n) = O(\frac{D_{SV_{\text{par}}}}{2}) + O(k|V|) + O(\frac{D^2_{SV_{\text{par}}}}{4}\overline{|Q|}|V|)$. To provide a comparison of the time complexities to compute the syntagmatic and paradigmatic measures, the following values are used, $D_{SV} = 20$, $k = 30$ and $\overline{|Q|} = 3$. These values are taken from those set or observed in the short query experiments carried out in Chapter 7. This results in the time complexity of each

measure being:

$$T_{\text{syn}}(n) = O(30|V|),$$

$$T_{\text{par}}(n) = O(100 \times 3 \times |V|) = O(300|V|),$$

where $T_{\text{syn}}(n)$ would be the same as the time complexity of RM3. This analysis demonstrates that for short queries, using TQE to include paradigmatic information within the query expansion process is likely to take 10 times longer than estimates using syntagmatic information alone. However, as the length of the queries increase the time complexity cost of the TQE approach will likely increase by a factor $\frac{10}{3}|Q|$.

6.4 Summary

This chapter formally applied the TE model to the task of query expansion within the relevance modelling framework. This technique was called *tensor query expansion* (TQE) and was designed to allow the mix of syntagmatic and paradigmatic information used to augment a query representation to be controlled so that a robust evaluation of the benefits of each can be made.

The computational efficiency of TQE was shown to be only slightly more than that of the unigram relevance model. This addresses one of the weaknesses traditionally associated with using dependency based approaches whilst at the same time preserving the improvements in effectiveness that term associations can provide.

The following chapter outlines a number of experiments used to evaluate the retrieval effectiveness of the TQE approach, along with the influence of syntagmatic and paradigmatic information in achieving this result.

Chapter 7

Evaluation of the Tensor Query Expansion Approach

A major premise behind using the TE model within the query expansion process stems from the fact that existing approaches primarily use syntagmatic information, and hence employ only half the associations reported to give rise to word meanings. Therefore, it was hypothesised that accessing information about both syntagmatic and paradigmatic information within the query expansion process may more effectively augment the query representations resulting in improved retrieval effectiveness.

Chapter 6 developed the *tensor query expansion* (TQE) approach that formally places the TE model within the relevance modelling framework. This chapter details a number of ad hoc retrieval experiments aimed at evaluating the benefits of using the TQE approach, with respect to strong benchmark relevance models, and provides a detailed examination of the improvements in retrieval effectiveness that may be gained by including information about paradigmatic associations.

These experiments represent different contexts in which the effectiveness of TQE and the importance of paradigmatic information can be evaluated, and comprise in summary:

1. **Short queries:** These experiments will use relatively short queries (often only 2 or 3 words in length), to simulate simulate the context often found within traditional web search engines.

   Due to the impact of insufficient sample data on the reliability of estimating paradigmatic associations, detailed in Section 4.4.2, it is hypothesised that similar issues relating to

insufficient sample data may be observed when dealing with short queries.

2. **Verbose queries:** These experiments will use relatively long queries, generally greater than 10 words in length. The long queries, also termed verbose queries, often form sentences seen within natural language, and are commonly found when performing question-answer tasks. Therefore, the results of these experiments will not only provide insight into the benefit of using syntagmatic and paradigmatic information to expand long queries, but also may provide insight into the potential value of using TQE within a question-answering context. Given the growing robustness of speech recognition systems and the increased prevalence of query suggestion functionality in search engines, it is expected that the use of verbose queries will be a growing trend in information retrieval research.

   It is hypothesised that for verbose queries the amount of data from which to model paradigmatic associations will be increased, and hence paradigmatic information may be able to play a more significant role in boosting retrieval effectiveness.

3. **2012 TREC WebTack:** This experiment will involve comparing the retrieval effectiveness of a system using TQE against other particpants within the 2012 TREC Web track, and a strong baseline system. TREC stands for ***T**xt **RE**trieval **C**onference*, and is a yearly event which provides researchers and industry an opportunity to pit their wits (and models) against each other on various information retrieval tasks[1]. There are very few limitations on what participants can use within their systems, with some teams utilising ontological resources and proprietary data such as web clickthrough logs. TREC has also been responsible for collating well accepted data sets, including document collections, queries and relevance judgements that allow robust evaluation between models to be performed. These data sets will be used within all three experimental settings in this chapter and are detailed in the following sections.

   It is hypothesised that by using a strong document model to form the set of pseudo relevant documents, more effective statistics should be available for modelling syntagmatic and paradigmatic associations, and hence further improvements in retrieval effectiveness should be achieved.

The first two experiments (i.e., for short and verbose queries) are aimed at investigating the role of each type of word association in helping improve retrieval effectiveness, and hence

[1] `http://trec.nist.gov`

fixes all free model parameters in TQE, except for the mix of syntagmatic and paradigmatic associations. The *TREC Web track* experiment on the other hand, evaluates the retrieval effectiveness of the TQE approach when all free model parameters are tuned. This difference in scientific method means that different types of findings are likely to be drawn. Therefore, the experimental results for the short and verbose query experiments have been separated out from those reported in the TREC Web track experiment.

7.1 Short and Verbose Query Experiments

7.1.1 Experimental Setup

Data Sets

Evaluation of all models was performed on the TREC data sets outlined in Table 7.1. All collections were stopped with the default 418 words Lemur stop list and stemmed using a Porter stemmer [Porter, 1997][2]. The experiments in this research were carried out using the Lemur Toolkit[3]. The Lemur implementation of the original positional relevance model is made available online by the original authors [4].

Queries

The queries used within the short and verbose experiments involve the title and description components of the TREC topics, respectively. The average length of title and descriptions for each data set are shown in Table 7.1 along with the standard deviation of each set of queries, which provides an idea of the range of query lengths for each data set[5].

Baseline and Benchmark Models

TQE was evaluated on an ad hoc retrieval task using pseudo relevance feedback, also known as blind feedback. The TQE approach, in Equation (6.8), was compared to a baseline unigram language model (i.e., with no pseudo relevance feedback and hence is denoted as **noFB**;

[2]The Clueweb document index used in these experiments was produced using a Krovetz stemmer.

[3]The Lemur toolkit for language modelling and information retrieval: http://www.lemurproject.org

[4]http://sifaka.cs.uiuc.edu/ ylv2/pub/prm/prm.htm

[5]Topics 1-50 in the Clueweb data set were not used as their relevance judgments were produced for the estimated AP metric [Yilmaz et al., 2008], which is not conceptually equivalent to those used for the MAP metrics.

	Description	# Docs	Topics	title $\overline{\lvert q \rvert}$	description $\overline{\lvert q \rvert}$	$\overline{\lvert D \rvert}$
WSJ	Wall Street Journal 87-92 off TREC Disks 1,2	173,252	1-200	4.8 (3)	19 (7.6)	468
AP	Assoc. Press 88-90 off TREC Disks 1,2,3	242,918	1-200	4.8 (3)	19 (7.6)	494
ROB	Robust 2004 data TREC Disks 4,5 -CR	528,155	301-450 601-700	2.6 (0.7)	16 (5.5)	561
G2	2004 crawl of .gov domain	25,205,179	701-850	2.28 (0.87)	11 (4.1)	1,721
CW	Clueweb09 Category B	50,220,423	Web Track 51-150	2.72 (1.38)	9 (3.3)	804

Table 7.1: Overview of TREC collections and topics. $\overline{|q|}$ represents the average length of the queries, the value in brackets is the standard deviation of the query lengths, and $\overline{|D|}$ is the average document length.

Section 5.2.2), a benchmark unigram relevance model (**RM3**; Section 5.3.2) and a positional relevance model using iid sampling (**PRM**; Section 5.3.2).

RM3 was chosen as a benchmark model primarily because it is a formal approach that fits within the language modelling framework, is robust [Lv and Zhai, 2009a] and has been used heavily as a benchmark for past query expansion research [Lv and Zhai, 2010, Metzler and Croft, 2007]. Even though the unigram relevance model does not explicitly model term dependencies, it was shown in Section 6.2.1 to effectively model syntagmatic associations with the query terms, and hence RM3's estimation technique was chosen as the TQE's syntagmatic feature. This decision was seen as an effective way to control the influence of paradigmatic information on retrieval effectiveness. This is because, if all other TQE and RM3 model parameters, except the mix of syntagmatic and paradigmatic information in TQE (i.e., γ in Equation (6.8)) are fixed, then any differences in retrieval effectiveness between TQE and RM3 can reliably be attributed to the influence of paradigmatic information.

A query expansion approach that explicitly models term dependencies was also chosen as a benchmark model. The choice was primarily between LCE and PRM, as these have been shown to significantly outperform RM3 [Lv and Zhai, 2010, Metzler and Croft, 2007]. PRM was chosen as it too fits within the relevance modelling framework, unlike LCE. This means that the set of pseudo relevant documents used by RM3, PRM and TQE for each query will be the same, as they will all use the same unigram language model for document ranking. PRM also has less free parameters than LCE. This means that improvements in retrieval effectiveness are less likely due to the increased degrees of freedom within the model [Metzler and Zaragoza, 2009].

Even though the reported retrieval effectiveness results published for LCE may suggest it provides superior retrieval effectiveness over PRM, it is important to recall two things; most importantly (i) that LCE is based on the MRF document ranking model while PRM is based on the unigram language model, and (ii) comparison of models based on published results is unreliable given experimental set up differences. Therefore, no comparative claim about the improved retrieval effectiveness of LCE over PRM, or *vice versa*, can be reliably made. It is also worth noting, that even though LCE and TQE both use a Markov random field to formalise the combination of features, this is merely a mathematical tool, and has little relevance to the types of linguistic associations being modelled within each approach, or the appropriateness of LCE as a benchmark model. For these reasons, the choice to use PRM over LCE as a second

benchmark approach appears justified.

It is acknowledged that PRM has not been evaluated on verbose queries in the past. However, PRM has significantly outperformed RM3 for short queries on large document collections [Lv and Zhai, 2010]. Therefore, the robust evaluation of PRM on smaller document collections for short queries and also verbose queries on all document collections constitutes a tangential contribution of this research.

Parameter Training

The baseline unigram language model that underpins all three relevance models being evaluated was done so using the Lemur default parameters. It has been acknowledged that increasing the number of parameters (degrees of freedom) within a model may in itself cause improvements in retrieval effectiveness [Metzler and Zaragoza, 2009]. However, to avoid the criticism that any model performs better due to an increased number of parameters and to control for the influence of paradigmatic information on retrieval effectiveness all common model variables in these two experiments were fixed. To this end, all expansion approaches were evaluated using 30 feedback documents, 30 expansion terms and a feedback interpolation co-efficient α=0.5 in Equation (5.22).

Even though it is common to fix one or more of these pseudo relevance feedback parameters [Bendersky et al., 2011b, Lafferty and Zhai, 2001, Lv and Zhai, 2010], it is acknowledged that the success of query expansion has been shown to be sensitive to the number of (pseudo) relevant documents and expansion terms [Billerbeck and Zobel, 2004, Ogilvie et al., 2009]. However, if the models under evaluation can produce significant improvements in retrieval effectiveness over the baseline unigram language model when these parameters are fixed, then it follows that greater improvements could be achieved if they were tuned.

Further support for the experimental choice of fixing all common query expansion parameters can be gained by noting that RM3 and TQE effectively use the same syntagmatic information. This is due to the syntagmatic feature in TQE using the same estimation technique as RM3. Therefore, any variations in retrieval effectiveness that could be achieved by tuning the common relevance model parameters, including the number of feedback documents, number of expansion terms and the feedback interpolation co-efficient (α in Equation (5.22)), would likely be seen in both. The claim that RM3 and TQE effectively use the same syntagmatic information

is also verified through comparison of expansion terms produced by TQE's syntagmatic feature and RM3, as demonstrated in Section 7.1.3.

For each of the query expansion techniques the free model parameters were trained using 3-fold cross validation on the MAP metric. This includes training the Dirichlet smoothing parameter, μ in the unigram relevance model of Equation (5.15). The free parameters trained for the positional relevance model included both σ and λ in Equation (5.26). For the TQE approach, the only free parameter was γ in Equation (6.8).

7.1.2 Experimental Results for Short Queries

Traditional web search often involves users entering very short, two or three word queries. To evaluate the impacts of including information about syntagmatic and paradigmatic associations to augment these short query representations within the information retrieval process an ad hoc retrieval experiment was carried out on the data sets and their topic titles outlined in Table 7.1. The mean average precision (MAP) and precision at 20 (P@20) for the top ranked 1000 documents for all models are reported in Table 7.2. The significance of the results was evaluated using a one sided paired t-test with $\alpha = 0.5$.

The results suggest that for short queries, the TQE approach can provide significant improvements over the baseline (noFB) and generally better effectiveness over the benchmark models on all data sets, except for CW. It appears RM3 and PRM are unable to achieve consistently significant retrieval effectiveness over the baseline. This is not consistent with previous results [Lavrenko and Croft, 2001, Lv and Zhai, 2010], but probably arises from the fixing of many free parameters in this experiment, including the number of feedback documents and expansion terms, to which query expansion approaches have been shown to be quite sensitive [Billerbeck and Zobel, 2004, Collins-Thompson, 2009, Ogilvie et al., 2009].

To confirm this, research by Symonds et al. [2011b] comparing the retrieval effectiveness of the RM3 and TQE approaches on short queries when all feedback parameters are tuned showed that TQE and RM3 provide significant improvements over the unigram language model (noFB) and TQE provides significant improvements over RM3. This confirms results previously reported using the RM3 approach, and supports the reasoning outlined in the previous paragraph.

The results (Table 7.2) do not negate our ability to rigorously evaluate the influence of syntagmatic and paradigmatic information within the query expansion process on retrieval

	Metric	noFB	RM3	PRM	TQE
WSJ	MAP	.2686	$.3089^{np}$ (15%)	$.3061^{n}$ (13.9%)	$\mathbf{.3090^{np}}$ (15%)
	P@20	.4074	$.4423^{n}$ (8.6%)	$.4413^{n}$ (8.3%)	$\mathbf{.4434^{n}}$ (8.8%)
AP	MAP	.1793	$.2144^{n}$ (19.6%)	$.2131^{n}$ (18.8%)	$\mathbf{.2145^{n}}$ (19.6%)
	P@20	.2300	$.2723^{n}$ (18.4%)	$.2788^{n}$ (22%)	$\mathbf{.2825^{n}}$ (22.8%)
ROB	MAP	.2500	$.2700^{n}$ (8%)	$.2707^{n}$ (8.3%)	$\mathbf{.2783^{nrp}}$ (11.3%)
	P@20	.3558	$.3688^{n}$ (3.7%)	.3639 (2.2%)	$\mathbf{.3741^{nrp}}$ (5.1%)
G2	MAP	.2941	$.3049^{n}$ (3.6%)	$.3069^{n}$ (4.3%)	$\mathbf{.3085^{n}}$ (4.9%)
	P@20	.5050	.5013 (-0.7%)	.5078 (0.5%)	$\mathbf{.5179^{nr}}$ (2.5%)
CW	MAP	.0768	.0778 (1.3%)	**.0822** (7.1%)	.0796 (3.7%)
	P@20	.1872	.1995 (6.5%)	**.2031** (8.4%)	.1995 (6.5%)

Table 7.2: Ad hoc retrieval results on short queries for the unigram language model (noFB), unigram relevance model (RM3), positional relevance model (PRM), and TQE (TQE). Statistically significant results ($p < 0.05$) are indicated by superscripts using the first letter of the baseline over which significant improvement was achieved (n=noFB, r=RM3, p=PRM, t=TQE). Bold indicates the best result for each dataset and metric. % improvement over noFB shown in parentheses.

effectiveness, as both RM3 and TQE's syntagmatic feature use the same estimation technique. It does however demonstrate that these results are likely to be very conservative indications of each model's possible retrieval effectiveness.

To gain insight into how the retrieval effectiveness of TQE compares on a *per query basis* to RM3 and PRM, a robustness analysis is presented.

Robustness

Robustness includes considering the ranges of relative increase/decrease in average precision and the number of queries that were improved/degraded, with respect to the baseline unigram language model (noFB). The graphs in Figure 7.1 and Figure 7.2 illustrate the relative

increase/decrease of average precision scores for the RM3, PRM and TQE approaches when evaluated on the ROB and G2 data sets, respectively.

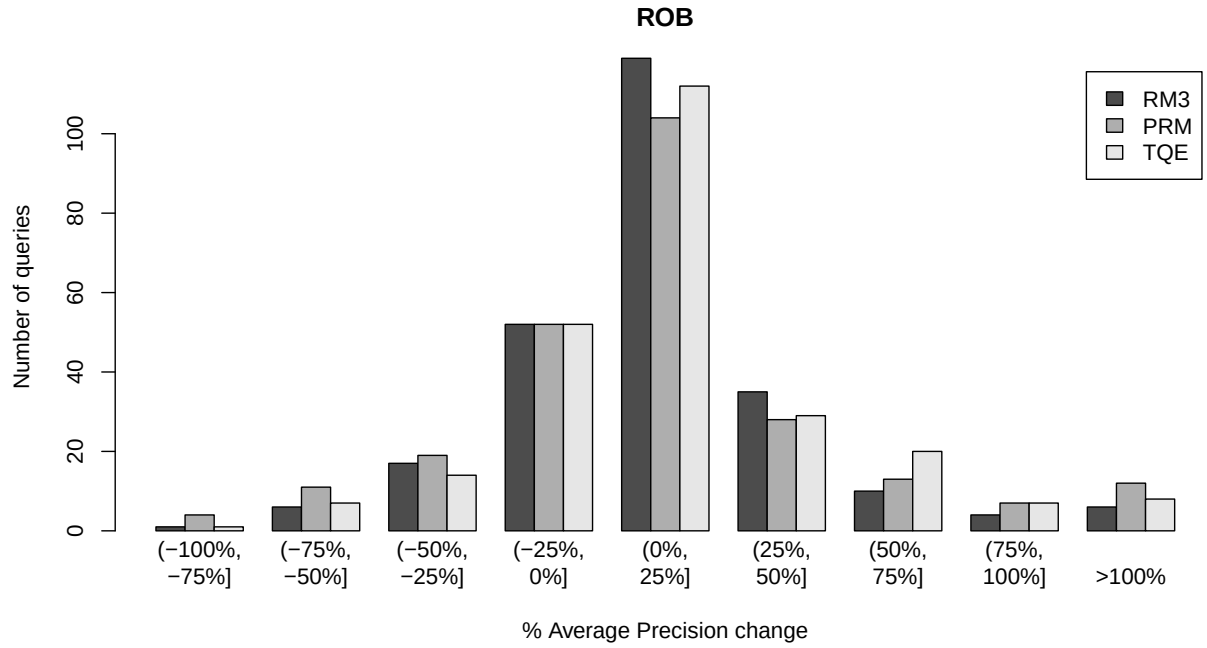

Figure 7.1: Robustness comparison of RM3, PRM and TQE on the ROB data set for short queries.

From looking at the distributions in Figure 7.1 and Figure 7.2, it may be argued that there is very little difference in robustness between the approaches, with TQE being slightly more robust than RM3 and PRM on the ROB data set, and RM3 being slightly more robust on the G2 dat set.

This inconsistent robustness result for TQE is an interesting result given TQE achieves a higher MAP score than RM3 and PRM on both the ROB and G2 data sets (refer to Table 7.2). This seemingly conflicting result likely indicates that when TQE *improves* the retrieval effectiveness over the baseline (noFB) it does so by a much greater percentage than RM3 and PRM. Alternatively, it may also be that when TQE hurts the retrieval effectiveness when compared to the baseline it does so by a much smaller percentage than RM3 and PRM.

If we consider the first possibility, then these larger improvements may relate to the usefulness of different linguistic associations for a given query. For example, a query that contains words that are predisposed to vocabulary mismatch, like TREC Topic 191 from the AP data set: *Efforts to improve U.S. schooling*, may benefit more from paradigmatic information, which may

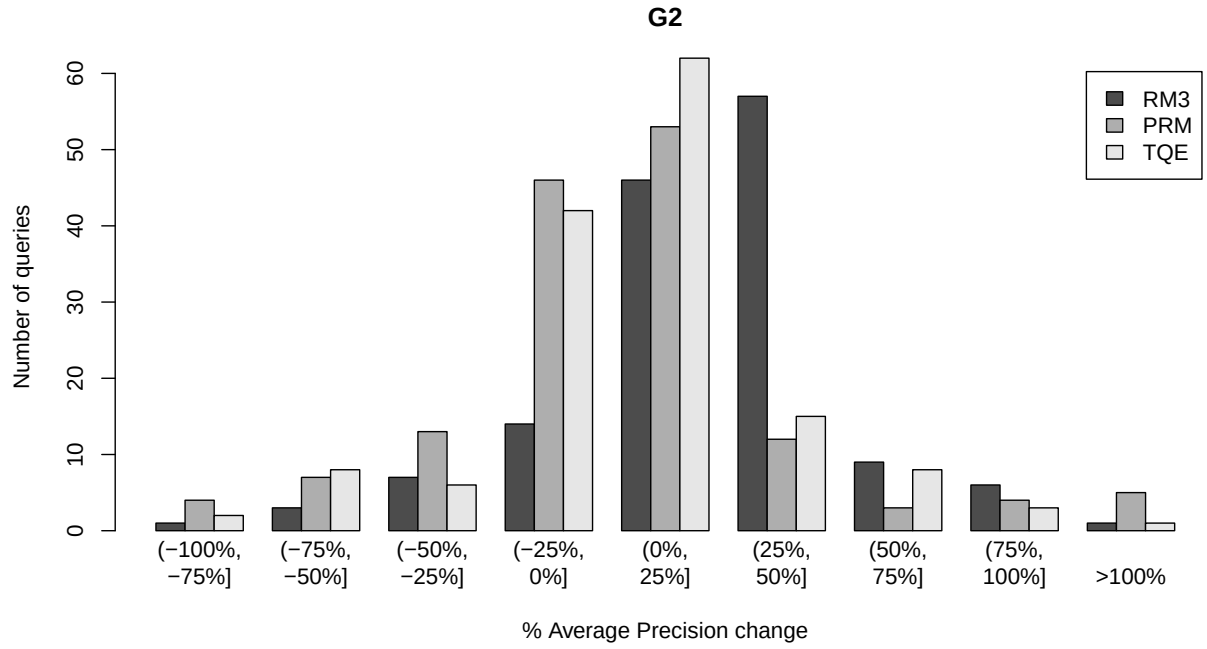

Figure 7.2: Robustness comparison of RM3, PRM and TQE on the G2 data set for short queries.

expand the query using words like: *attempts, research, enhance, lift, united states, american, teaching, academic results*.

In fact, the best retrieval effectiveness (MAP $= 0.334$) for this query (TREC Topic 191) on the AP data set was achieved when purely paradigmatic information ($\gamma = 1$) was used within the TQE approach. When compared to the effectiveness (MAP $= 0.211$) achieved when using the best γ that returned the best average effectiveness, some linguistic motivation for predicting γ based on the query terms could be argued. The validity of performing an oracle analysis based on the possible link between query terms and the best mix of syntagmatic and paradigmatic information for that query is discussed further in Section 7.3.

To gauge how many queries in each data set resulted in an improved average precision (AP) score over the baseline for each query expansion approach, the improvement (IMP) metric can be used [Xu et al., 2009]. IMP scores for RM3, PRM and TQE on all data sets using short queries are reported in Table 7.3. These highlight the discrepancy that even though RM3 improved the AP score over the baseline on more queries than TQE on a number of data set, TQE still achieves a higher MAP score on these data sets. This result suggests that when TQE improves the retrieval effectiveness for a given query, it does so by a much larger percentage than RM3. To determine whether this is a result of the use of paradigmatic information a

	WSJ (/200)	AP (/200)	ROB (/250)	G2 (/150)	CW (/100)
RM3	151	149	173	77	54
PRM	147	106	163	77	55
TQE	149	147	175	89	54

Table 7.3: IMP for RM3, PRM and TQE on each of the data sets using short queries. IMP indicates the number of queries that achieved an improved MAP score over the baseline (noFB) MAP. The number of queries in the data set are shown in brackets next to the data set name.

parameter sensitivity analysis is required.

Parameter sensitivity

To understand the role that syntagmatic and paradigmatic information plays in achieving the reported retrieval effectiveness of TQE a parameter sensitivity analysis was performed. This analysis, shown in Table 7.4, displays the value of γ in Equation (6.8) that provides the best retrieval effectiveness for each data set using short queries[6].

	WSJ	AP	ROB	G2	CW
γ that produced the maximum MAP	0	0	0.1	0.2	0.4
γ that produced the maximum P@20	0.2	0 / 0.4	0.2	0.8	0.4

Table 7.4: Parameter sensitivity analysis showing the value of γ in TQE that produces the best MAP and P@20 for the TREC collections using short queries.

The results (Table 7.4) suggest that the information about paradigmatic associations do not play a major role in providing the best MAP or P@20 scores on small newswire document collections like WSJ, AP and ROB. However, the role of paradigmatic information appears to become more important when searching large web collections, such as those used within the G2 and CW data sets.

It is hypothesised that this reduced reliance on paradigmatic information for smaller newswire

[6]This analysis does not use a train/test split

collections may be due to the superior ability to model syntagmatic associations in these collections, as compared to modelling paradigmatic associations. This idea stems from the increased likelihood that query terms and effective expansion terms will co-occur within the same document for small collections that likely have little noise. However, for larger, noisy web collections the possibility that effective expansion terms co-occur within the same document as the query terms is reduced, and hence syntagmatic associations are less effectively modelled. In these cases, the modelling of paradigmatic associations are likely to be as reliable, if not more, than syntagmatic associations, as shown by the increased dependence on paradigmatic information for the G2 and CW data sets (Table 7.4).

This parameter sensitivity analysis appears to have highlighted a possible inconsistency in the experimental results, in that the IMP scores reported earlier for the WSJ and AP data sets suggested that RM3 improves the MAP score over the baseline for two more queries than TQE does. However, Table 7.4 indicates TQE is using only the syntagmatic information (i.e., $\gamma = 0$, and hence TQE should be equivalent to RM3) for these data sets. Therefore, it would be expected that the IMP scores for RM3 and TQE would be identical on these data sets. However, it is important to remember that the parameter sensitivity results (Table 7.4) are produced without using a train/test split based on 3-fold cross validation. This means the reported results in Table 7.2 and Table 7.3 are created from a cross-validation process and hence the free parameter value may not be the same on all queries. This slight aberration highlights the importance of locking down as many free parameters as possible within the models being evaluated so that the confidence with which the influence of paradigmatic information on retrieval effectiveness can be assessed.

Conclusion for Experiments on Short Queries

An important finding from this experiment relates to the fact that for short queries, the inclusion of paradigmatic information *does not* consistently enable TQE to significantly outperform query expansion techniques like RM3 and PRM that predominantly use syntagmatic information. It is hypothesised that this may be due to the reduced sample sizes associated with only a small number of query terms from which paradigmatic associations need to be modelled. If this were the case then we can hypothesise that using TQE on longer queries should increase the dependence on paradigmatic information in achieving the best retrieval effectiveness, and hopefully also provide more robust improvements in retrieval effectiveness. This hypothesis

will be tested as a part of the experiment investigating the performance of TQE on verbose queries.

7.1.3 Experimental Results for Verbose Queries

Long queries make up a smaller yet important proportion of web queries submitted to search engines, and are common in collaborative question answering (QA) [Balasubramanian et al., 2010, Bendersky and Croft, 2009, Huston and Croft, 2010]. A recent report produced by the information retrieval community also identified *conversational answer retrieval* as one of the six most important topics for future information retrieval research [Allan et al., 2012].

	Metric	noFB	RM3	PRM	TQE
WSJ	MAP	.2121	$.2682^{n}$ (26.4%)	$.2589^{n}$ (22.1%)	$\mathbf{.2865^{nrp}}$ (35.0%)
	P@20	.3480	$.3891^{n}$ (11.8%)	$.3795^{n}$ (9.1%)	$\mathbf{.4149^{nrp}}$ (19.2%)
AP	MAP	.1511	$.1991^{n}$ (31.8%)	$.1861^{n}$ (23.2%)	$\mathbf{.2056^{nrp}}$ (36.1%)
	P@20	.2300	$.2600^{n}$ (13.0%)	.2458 (6.8%)	$\mathbf{.2738^{nrp}}$ (19.0%)
ROB	MAP	.2491	$.2643^{n}$ (6.1%)	$.2704^{n}$ (8.5%)	$\mathbf{.2869^{nrp}}$ (15.1%)
	P@20	.3373	.3414 (1.2%)	$.3504^{nr}$ (3.9%)	$\mathbf{.3650^{nrp}}$ (9.1%)
G2	MAP	.2466	$.2571^{n}$ (4.3%)	$.2583^{n}$ (4.8%)	$\mathbf{.2719^{nrp}}$ (10.3%)
	P@20	.4594	.4620 (0.6%)	.4732 (1.1%)	$\mathbf{.4842^{nrp}}$ (5.4%)
CW	MAP	.0530	.0558 (5.2%)	$\mathbf{.0614^{n}}$ (16.3%)	.0574 (8.3%)
	P@20	.1561	.1566 (0.3%)	$\mathbf{.1724^{n}}$ (10.5%)	.1607 (2.9%)

Table 7.5: Ad hoc retrieval results for verbose queries, for the unigram language model (noFB), unigram relevance model (RM3), positional relevance model (PRM), and TQE (TQE). Statistically significant results ($p < 0.05$) are indicated by superscripts using the first letter of the baseline over which significant improvement was achieved (n=noFB, r=RM3, p=PRM, t=TQE). Bold indicates the best result for each dataset and metric. Brackets indicate percent improvement over noFB.

The mean average precision (MAP) and precision at 20 (P@20) for the top ranked 1000 documents for all models evaluated are reported in Table 7.5. The significance of the results

were evaluated using a one sided paired t-test with $\alpha = 0.5$. The results suggest that TQE can provide significant improvement over the baseline and benchmark models on all data sets, except for CW. These results also indicate that the improvements in retrieval effectiveness of RM3 and PRM are not always significantly better than the baseline (noFB). However, this is likely due to the fixing of all other pseudo relevance feedback parameters, including the number of feedback documents and expansion terms. Therefore, these results can be considered a conservative indication of performance. Given this, any the following should be restricted to better understanding the influence of paradigmatic information on information retrieval effectiveness, not making generalised claims regarding the comparative performance of the models, especially between TQE and PRM.

The graph in Figure 7.3 illustrates the ability of the TQE paradigmatic feature to improve retrieval effectiveness as the average length of queries in a data set increases. This graph also illustrates how as the average length of the queries in a data set reduces, to a minimum average of 9 on the CW data set, the retrieval effectiveness advantage of TQE over PRM appears to also reduce, and even falls behind PRM on the CW data set.

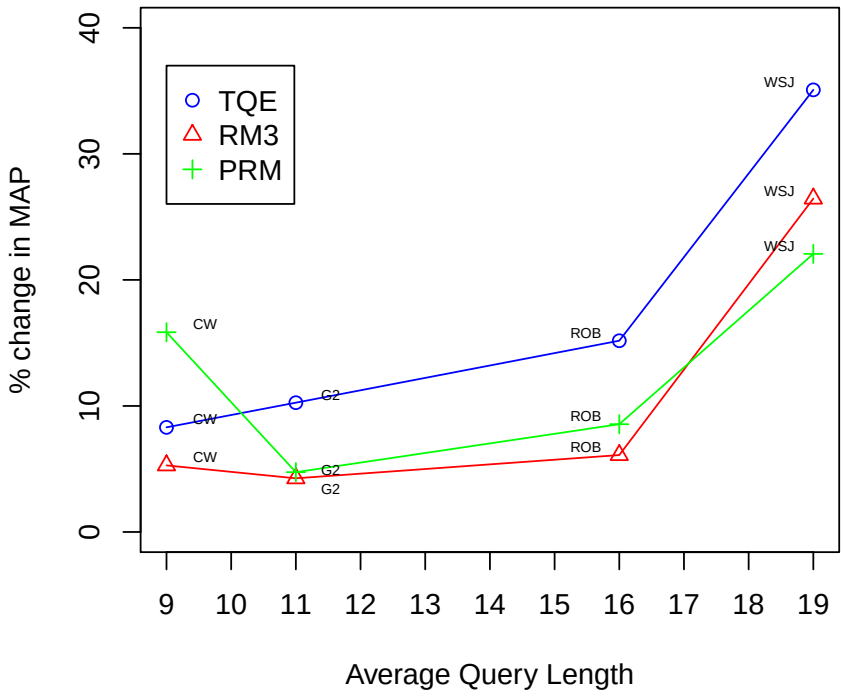

Figure 7.3: Percent improvement in MAP of RM3, PRM and TQE over the unigram language model (noFB) for the average query lengths of the data set descriptions listed in Table 7.5.

Based on the results from the previous experiment on short queries, it is hypothesised that this drop in performance when compared to the benchmark models on the CW data set may be due to the average query length being too short to allow effective paradigmatic associations to be modelled within TQE's paradigmatic feature. To gain a better insight into how much of the drop off in retrieval effectiveness can be attributed to the query length, the effectiveness of TQE for queries with a minimum length of 11 in the CW data set is evaluated.

	Metric	noFB	RM3	PRM	TQE
CW_v	MAP	.0681	$.0816^{n}$	$.0827^{n}$	$\mathbf{.0882^{nrp}}$
			(19.7%)	(21.4%)	(29.4%)
	P@20	.2267	.2417	$.2423^{n}$	$\mathbf{.2500^{nrp}}$
			(6.6%)	(6.9%)	(10.3%)

Table 7.6: Ad hoc retrieval results for verbose queries ($|q| > 10$) on the CW_v data set, for the unigram language model (noFB), unigram relevance model (RM3), positional relevance model (PRM), and TQE (TQE). Statistically significant results ($p < 0.05$) are indicated by superscripts using the first letter of the baseline over which significant improvement was achieved (n=noFB, r=RM3, p=PRM, t=TQE). Bold indicates the best result for each dataset and metric. Brackets indicate percent improvement over noFB.

Our choice of minimum query length was based on providing a balance between previous research, including work by Bendersky and Croft [2009] where verbose queries were defined as having a length greater than 12, and choosing a query length which would ensure sufficient data samples for a meaningful analysis. For the CW data set, the number of topics with $|q| > 10$ was 30, with $|q| > 11$ was 16 and with $|q| > 12$ was 11. Therefore, queries with length greater than 10 were chosen for this evaluation, as indicated by the CW_v data set in Table 7.1. The retrieval effectiveness results on the CW_v data set are shown in Table 7.6 and demonstrate that TQE achieves significant improvement in retrieval effectiveness over the baseline and benchmark models for this data set.

To gain a better insight into how the retrieval effectiveness of the TQE approach compares on a *per query basis* to that of RM3 and PRM, a robustness analysis is required.

Robustness

The graphs in Figure 7.4 and Figure 7.5 illustrates the relative increase/decrease of average precision scores for the RM3, PRM and TQE approaches over the unigram language model (noFB) when evaluated on the ROB and G2 data sets, respectively.

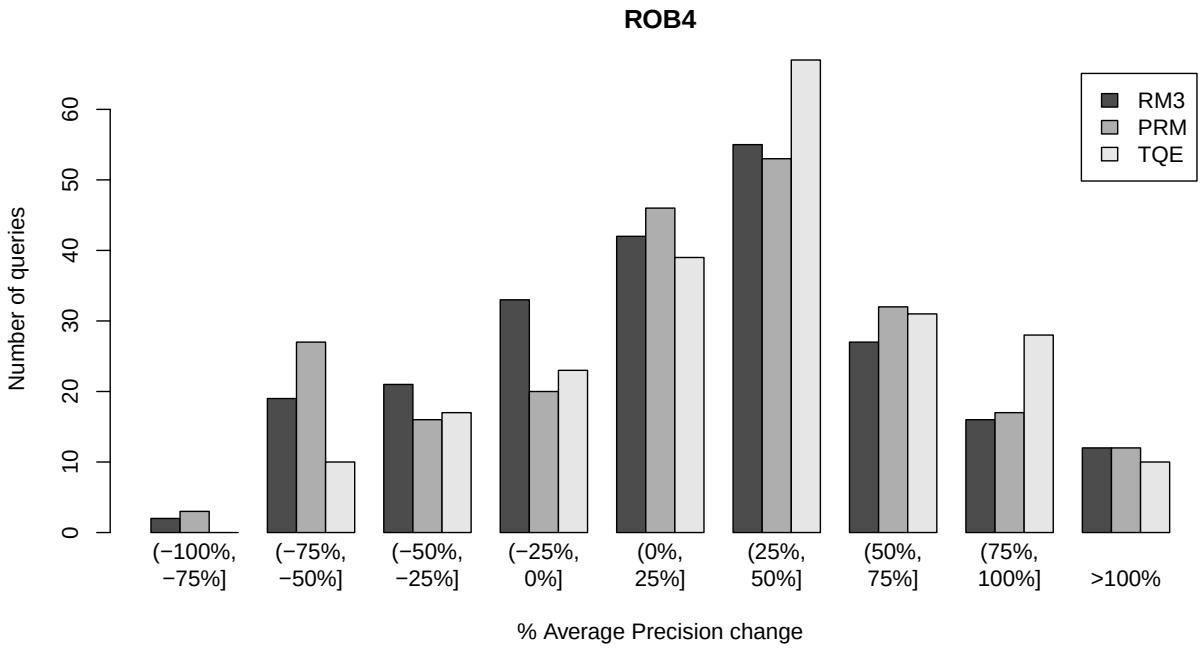

Figure 7.4: Robustness comparison of RM3, PRM and TQE on the ROB data sets for verbose queries.

This analysis suggests that TQE provides more consistent improvements over the baseline language model (noFB) than the unigram relevance model (RM3) and positional relevance model (PRM). The graphs for the other data sets were omitted for space reasons, however, a similar result was observed. To more easily see how many queries on each data set achieved an improved MAP score over the baseline (noFB) the IMP score for RM3, PRM and TQE on all data sets is reported in Table 7.7.

For verbose queries, as opposed to shorter queries, the IMP scores suggest that the TQE approach does consistently improve the retrieval effectiveness of more queries than RM3 and PRM when compared to the baseline, using the feedback parameter outlined in Section 7.1.1.

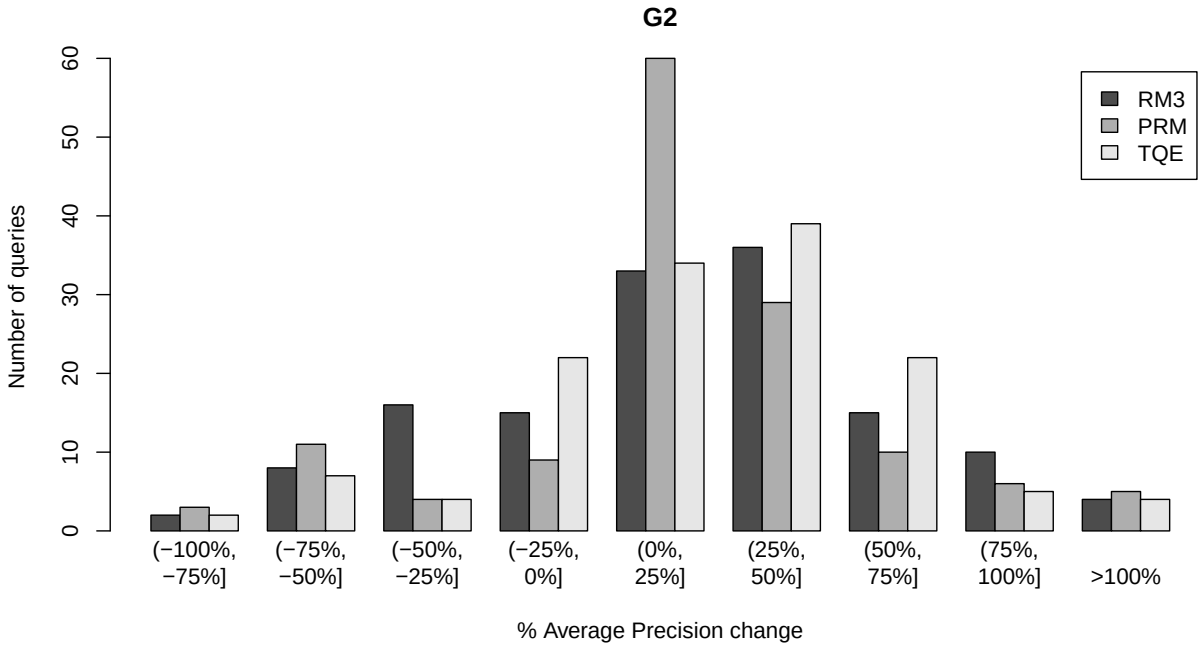

Figure 7.5: Robustness comparison of RM3, PRM and TQE on the G2 data sets for verbose queries.

Parameter Sensitivity

To understand the role that syntagmatic and paradigmatic information plays in achieving the reported retrieval effectiveness of TQE, a parameter sensitivity analysis was performed. This analysis, shown in Table 7.8 displays the value of γ in Equation (6.8) that provides the best retrieval effectiveness[7].

The results (Table 7.8) suggest that for verbose queries information about paradigmatic associations are consistently contributing to the ability to achieve maximum retrieval effectiveness within the TQE approach. This result appears to support the hypothesis, raised in Section 7.1.2, that short queries do not provide enough statistical evidence to effectively model paradigmatic associations.

The increased reliance on the paradigmatic feature for the AP data set, when compared to the WSJ data set, which uses the same topics, may suggest that fewer of the initially retrieved documents from the AP collection were relevant and therefore less within document co-occurrences of query terms and effective expansion terms existed, leading to ineffective modelling of syntagmatic associations. This is supported by the relatively low MAP of the

[7]This analysis does not use a train/test split

	WSJ (/200)	AP (/200)	ROB (/250)	G2 (/150)	CW (/100)	CW_v (/30)
RM3	117	114	132	73	44	14
PRM	112	121	133	59	40	16
TQE	144	141	160	80	49	19

Table 7.7: IMP for RM3, PRM and TQE on each of the data sets evaluated. IMP indicates the number of queries that achieved an improved MAP score over the baseline (noFB) MAP. The number of queries in the data set are shown in brackets next to the data set name.

	WSJ	AP	ROB	G2	CW	CW_v
γ that produced the maximum MAP	0.1	0.4	0.2	0.6	0.4	0.4
γ that produced the maximum P@20	0.2	0.5	0.3	0.6	0.2	0.5

Table 7.8: Parameter sensitivity, showing the value of γ in TQE that produces the maximum MAP and precision at 20 scores for the TREC collections.

unigram language model (noFB) on the AP data set when compared to the WSJ data set. This suggests that for more difficult queries, using information about paradigmatic associations may be more effective at improving retrieval effectiveness. This result also suggests that using word distributions found within more relevant documents may improve the effectiveness of the modelling within the TQE approach. This idea is tested further in the TREC Web track experiment outlined in Section 7.2.

Expansion Term Comparison

To confirm that the benchmark models use primarily syntagmatic information and that the use of paradigmatic information is unique to the TQE approach, a Jaccard coefficient and Spearman's rank correlation coefficient analysis on the set of expansion terms was performed. The Jaccard coefficient analysis will measure the average number of expansion terms that are common between two approaches. The Spearman's rank correlation coefficient is a finer grained analysis which measures, on a per query basis, how similar the overlap of two sets of expansion terms are with a third set.

		RM3	PRM	TQE	$s_{\text{syn}}(,)$
G2	PRM	.509 (20)	1 (30)		
	TQE	.725 (25)	.439 (18)	1 (30)	
	$s_{\text{syn}}(,)$	.780 (26)	.488 (20)	.643 (23)	1 (30)
	$s_{\text{par}}(,)$	.104 (6)	.108 (6)	.222 (10)	.103 (6)
CW	PRM	.634 (23)	1 (30)		
	TQE	.634 (23)	.528 (21)	1 (30)	
	$s_{\text{syn}}(,)$	.768 (26)	.598 (22)	.697 (25)	1 (30)
	$s_{\text{par}}(,)$	.131 (7)	.130 (7)	.238 (12)	.128 (7)

Table 7.9: Average Jaccard co-efficients for the sets of expansion terms produced on the G2 and CW data sets using verbose queries for the best performing RM3, PRM and TQE as well as the syntagmatic and paradigmatic features. The average number of expansion terms that overlap between approaches is shown in brackets.

When compared to RM3, the syntagmatic feature $s_{\text{syn}}(,)$ has a minimum Jaccard coefficient of 0.768 (Table 7.9). This means that on average at least 25 out of the 30 expansion terms suggested by $s_{\text{syn}}(,)$ are in common with those suggested by RM3. $s_{\text{syn}}(,)$ also has a minimum Jaccard coefficient of 0.488 with PRM. Indicating that on average at least 20 of the 30 expansion terms are in common.

As an example of the types of terms being produced by each of the query expansion techniques, Table 7.10 lists the top 10 query terms and estimates for TREC topic 148: *Find information about Martha Stewarts insider trading case;* on the Clueweb09 CategoryB document collection. It can be seen from this table that the top 10 expansion terms for RM3 are identical and in the same order as those produced by the TQE syntagmatic feature, $s_{\text{syn}}(,)$. The bottom row of Table 7.10 highlights the percent change in MAP over the baseline unigram language model produced by each approach.

To investigate the overlap for each topic, a per-topic Spearman's rank correlation coefficient (ρ) analysis, along the number of overlapping expansion terms on the $s_{\text{par}}(,)$ feature, was performed for the RM3, PRM and $s_{\text{syn}}(,)$ approaches. The resulting coefficients were,

PRM	RM3	TQE	$s_{\text{par}}()$	$s_{\text{syn}}()$
martha (.0842)	martha (.0510)	martha (.0295)	find (.0016)	martha (.0728)
stewart (.0686)	stewart (.0412)	stewart (.0233)	information (.0015)	stewart (.0563)
new (.0121)	insider (.0402)	insider (.0204)	trade (.0015)	insider (.0503)
com (.0081)	trade (.0131)	trade (.0075)	timeline (.0014)	trade (.0165)
site (.0081)	new (.00945)	new (.0046)	case (.0013)	new (.0115)
live (.0071)	com (.0058)	com (.0033)	stewart (.0013)	com (.0077)
insider (.0057)	site (.0053)	site (.0025)	theme (.0008)	site (.0058)
home (.0050)	home (.0046)	home (.0024)	lawyer (.0007)	home (0.0057)
official (.0049)	article (.0040)	article (.0021)	invest (.0007)	article (.0052)
photo (.0048)	stock (.0037)	information (.0020)	martha (.0007)	stock (.0047)
ΔMAP -54%	ΔMAP +18%	ΔMAP +16%	ΔMAP +23%	ΔMAP +16%

Table 7.10: Top 10 expansion terms and their estimates for TREC Web Track topic 148 (*Find information about Martha Stewarts insider trading case*) on the Clueweb09 CategoryB document collection for RM3, PRM, TQE and the paradigmatic and syntagmatic features.

$\rho_{(PAR:SYN,RM3)} = 0.941$, $\rho_{(PAR:SYN,PRM)} = 0.863$ and $\rho_{(PAR:RM3,PRM)} = 0.883$. This would indicate that the number of expansion terms overlapping on a per-topic basis for RM3, PRM and $s_{\text{syn}}(,)$ with the $s_{\text{par}}(,)$ feature are very similar.

This analysis demonstrates that both RM3 and PRM use very similar term dependencies to the TQE syntagmatic feature, $s_{\text{syn}}(,)$. Therefore, it is argued that the decision to lock down all other model parameters has meant that the primary reason for differences in retrieval effectiveness between TQE and the benchmark models is the inclusion of paradigmatic associations in the query expansion process.

Impact of Paradigmatic Information

The Jaccard co-efficients in Table 7.9 show that the expansion terms produced by PRM, RM3 and $s_{\text{syn}}(,)$ have little overlap with those produced by the TQE's paradigmatic feature, $s_{\text{par}}(,)$. For example, for RM3 and PRM, on average a maximum of 7 out of the 30 expansion terms are in common with those suggested by $s_{\text{par}}(,)$. This suggests that these approaches are likely using very little paradigmatic information when estimating conditional probabilities of expansion terms. Especially given the ability of a single term to be considered to have syntagmatic and paradigmatic associations with a query term. This means a term may be used as an expansion term in the RM3 and PRM approaches due to its syntagmatic association with a the query terms, but the same term may exist in the set of expansion terms for the paradigmatic reason as it also has strong paradigmatic associations with the query terms. For example, if *coffee* is a query term, then the word *tea* may be listed as an expansion term by the syntagmatic feature (or RM3 and PRM) as *we have tea or coffee* highlights a syntagmatic associations, but also by the paradigmatic feature, as *a cup of tea* or *a cup of coffee* highlights the paradigmatic association between *coffee* and *tea*.

A final interesting point raised by observing the syntactic class of the expansion terms produced by $s_{\text{par}}(,)$ across the CW data set, is that these paradigmatic associations do not often manifest as synonyms or antonyms (commonly adjectives). The $s_{\text{par}}(,)$ expansion terms appear more likely to be related verbs, like *trade-invest* in Table 7.10. This result is seen as an attractive feature of the TQE approach and may help explain why ontological based attempts at using paradigmatic information within the query expansion process have not been overly successful, like those using WordNet [Voorhees, 1994].

7.1.4 Conclusion

The experiment on verbose queries has demonstrated that using both syntagmatic and paradigmatic information to enhance query representations within the TQE approach can lead to significantly improved retrieval effectiveness over a baseline and two strong benchmark query expansion approaches. This result was achieved over a wide range of data sets, but needs to be considered in view of the parameter fixing steps outlined in Section 7.1.1.

Evaluation of the TQE approach on short queries appeared to suggest that these queries did not provide enough statistical data to make reliable estimates of paradigmatic associations. This

was supported by the finding that on verbose queries the reliance on paradigmatic information to achieve significant improvements in retrieval effectiveness was increased.

Experiments on both short and verbose queries suggested that the role of paradigmatic information to improve retrieval effectiveness increases as the size and likely noise within the document collections increases. This may be due to the reduced reliability of estimating syntagmatic associations within these noisy environments.

The second research question, posed in Chapter 1, asked if using a corpus-based model of word meaning to enhance query representations could provide improvements in retrieval effectiveness for current state-of-the-art systems. Therefore, to answer this question the TQE approach should also be evaluated when all free parameters in a (pseudo) relevance feedback setting are tuned. This evaluation will be performed within the setting of a highly regarded international information retrieval forum, so as to compare the retrieval effectiveness of our TQE submission against leading industry and academic systems.

7.2 The 2012 TREC Web Track

To evaluate how the TQE approach performs against state-of-the-art information retrieval systems a number of submissions were made in the 2012 TREC Web track, under the participant team name of QUT_Para [Symonds et al., 2013]. This forum provides an opportunity to pit the TQE approach against industry and research teams from all over the world. These teams have access to some impressive resources.

7.2.1 Experimental Setup

In this experiment the set of training documents used to build the TE model's vocabulary used within the TQE approach is based on the set of k top ranked pseudo relevant documents produced by a very strong baseline model. The baseline model is created using the following approach:

- The *ClueWeb09-Category B* documents are indexed using the 'indexing without spam'

approach [Zuccon et al., 2011]. Each query is then issued to the Google retrieval service[8] and the top 60 retrieved documents are filtered using the spam filtered ClueWeb09-Category B index[9]. On average, 13 out of the 60 top ranked Google documents existed in this index. This filtered list is then padded, to create a list of 10,000 documents, using the list of documents returned from a search on the spam filtered index using a unigram language model. These rankings form the baseline submission (**QUTParaBline**). The use of Google as the search engine for the top ranked results and the filtering of spam web pages are likely to translate into a strong baseline.

The processes underpinning the baseline (QUTparaBline) and TQE (QUTparaTQEg1) submissions are depicted in Figure 7.6 and Figure 7.7, respectively.

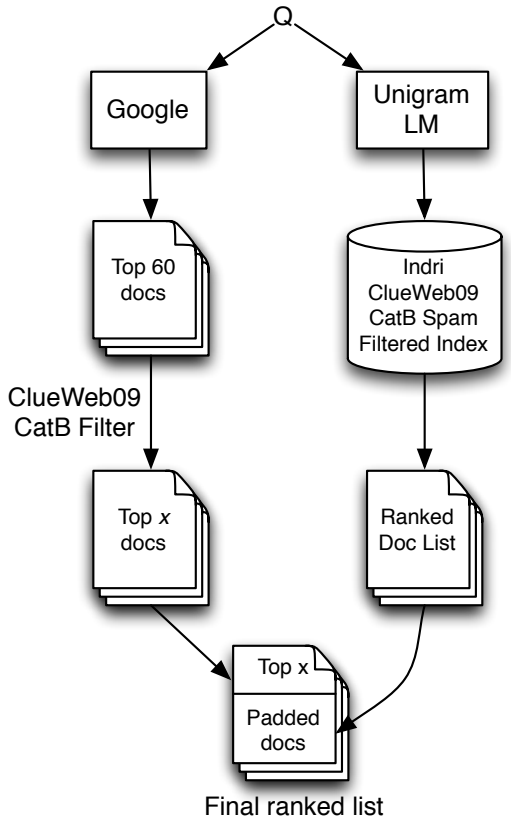

Figure 7.6: The baseline: QUTParaBline.

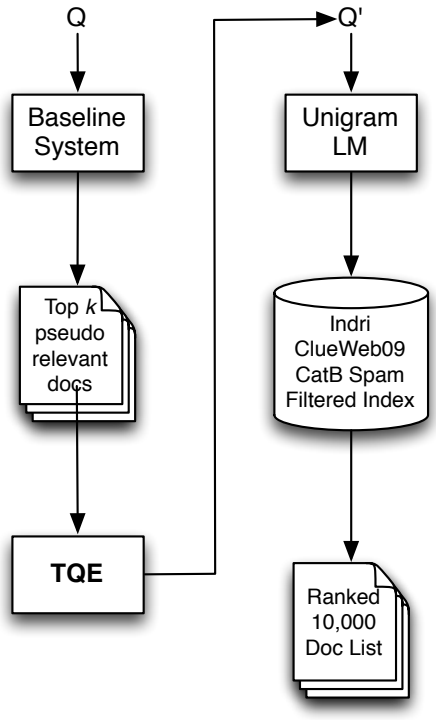

Figure 7.7: TQE: QUTParaTQEg1.

[8] http://www.google.com

[9] We had to limit the number of documents retrieved with Google to 60 because of Google's policies regarding the retrieval service at the time.

Training TQE

The data set used for this experiment is shown in Table 7.11. In this experiment all parameters in the pseudo relevance feedback setting were trained, these include the number of feedback documents (fbDocs), number of expansion terms (fbTerms), the mix of original and new query models (α) and the mix of syntagmatic and paradigmatic information (γ). Tuning of the QUTParaTQEg1 system parameters was achieved via training on ERR@20 metric using the TREC web Track data sets from 2010 and 2011. The test topics were those provided by TREC for the 2012 Web track. Participants only receive the topic titles for producing their submissions. Details regarding descriptions and further relevancy information are only provided after all submissions have been evaluated.

	Description	# Docs	Topics	title $\overline{\lvert q \rvert}$	description $\overline{\lvert q \rvert}$	$\overline{\lvert D \rvert}$
CW	Clueweb09 Category B	50,220,423	Web Track 51-200	2.72 (1.38)	9 (3.3)	804

Table 7.11: TREC collections and topics used creating the QUT_Para TREC submissions. $\overline{|q|}$ represents the average length of the queries, the value in brackets is the standard deviation of the query lengths, and $\overline{|D|}$ is the average document length.

The test parameter values for the QUTParaTQEg1 submission were *Number of feedback documents* equal to 19, *number of expansion terms* equal to 14, *original query weight* equal to 0.4 and *TE model mixing parameter* (γ) equal to 0.1. A value of $\gamma = 0.1$ again demonstrates that some combination of both syntagmatic and paradigmatic information provides optimal retrieval effectiveness (Section 7.1.3,Chapter 4).

7.2.2 Experimental Results

The document collection used in our experiments were based on a spam filtered version of the ClueWeb09 Category B corpus. The spam filtering was done using the Waterloo spam listing and a threshold of 0.45 Cormack et al. [2011]. The remaining documents were indexed using

the Indri toolkit [Strohman et al., 2005][10]. The Index was stopped using a standard INQUIRY stopword list [Allan et al., 2000] and stemmed using a Krovetz stemmer [Krovetz, 1993].

Comparison of the Runs

In this section we compare the results of the two runs submitted to the ad hoc task of the TREC 2012 Web Track. The two runs submitted represented a baseline (without query expansion) and a second using the TQE approach.

1. **QUTParaBline**: This run was produced by padding the ClueWeb09 Category B, spam filtered, top 60 Google results with the results returned by a unigram language model on the same spam filtered index (refer to Figure 7.6). This run forms the baseline.

2. **QUTParaTQEg1**: This run was produced by expanding the original TREC 2012 Web Track topics using TQE based on a set of k pseudo-relevant documents produced by the baseline model (refer to Figure 7.7).

Table 7.12 compares the retrieval effectiveness of these runs (QUTparaBline, QUTparaTQEg1) along with the average effectiveness of all 48 TREC Web track submissions on the ClueWeb09 CategoryB collection (MeanWT2012), and a baseline unigram language model (unigramLM).

These results suggest that expanding the query representations using TQE can provide significant improvements over the Google baseline on the binary metrics of MAP and P@20. No significant difference in retrieval effectiveness was noted on the graded metrics (ERR@20 and nDCG@20).

Graded metrics are those that base their effectiveness score on documents that are assigned a relevance judgement in a range, i.e., between 0 and 4. In addition, measures such that use graded judgements, such as ERR, often bias the scores for systems that return relevant documents toward the very top of the ranked list (i.e., in positions 1,2 and 3 say). This causes a heavy discounting to occur for relevant documents ranked lower in the list, as seen from the expression used to calculate ERR at rank k (Equation 5.4).

Given Google rankings are likely based on click through data and editorial choice, the QUTParaBline system is able to ensure relevant documents are ranked high in the returned list.

[10] Available at http://scourceforge.net/projects/lemur

	Graded Metrics		Binary Metrics	
	ERR@20	nDCG@20	P@20	MAP
unigramLM	0.160	0.112	0.254	.107
MeanWT2012	0.187	0.123	0.284^{u}	-
QUTparaBline	**0.290**^{um}	0.167^{um}	0.305^{um}	0.117^{u}
QUTparaTQEg1	0.249^{um}	**0.192**^{um}	**0.396**^{umb}	**0.158**^{ub}
	(-14.2%)	(+15%)	(+29.8%)	(+35%)

Table 7.12: Comparison of retrieval performance on TREC 2012 Web Track ad hoc retrieval task. The superscripts u, m, b and t indicate statistically significant differences (calculated using a paired t-test $p < 0.05$) over the unigram language model (unigramLM), the average performance of all TREC Web track participants (MeanWT2012), our baseline (QUTparaBline) and the TQE approach (QUTparaTQEg1), respectively. The best results for each evaluation measure appear in boldface. Brackets indicate the percentage change between QUTparaTQEg1 and QUTparaBline. Note that no value of MAP was provided for the average of all TREC 2012 Web Track submissions (MeanWT2012).

However, as the QUTParaTQEg1 system performs its final ranking using a unigram language model, which does not use such information, it is not surprising that the QUTParaTQEg1 model is unable to achieve significant improvements over QUTParaBline on the graded metrics ERR@20 and nDCG@20. In fact, it is surprising that QUTParaTQEg1 was even able to outperform QUTParaBline on nDCG@20.

Recall, the QUTParaTQEg1 system achieved significant improvements over QUTParaBline on the P@20 metric (Table 7.12). This indicates that many more relevant documents were returned in the top 20 by QUTParaTQEg1 than QUTParaBline. Given this result, significant improvements on graded metrics, such as ERR and nDCG, may be achievable if the final document ranking model used in QUTParaTQEg1 was enhanced to take into account ranking preferences.

Robustness

Robustness includes considering the ranges of relative increase/decrease in effectiveness and the number of queries that were improved/degraded, with respect to the baseline (MeanWT2012). The graph in Figure 7.8 illustrates the relative increase/decrease of P@20 scores for QUTParaBline and QUTParaTQEg1 when evaluated on the test topics (151-200) of the CW data set[11]. This graph suggests that the QUTParaTQEg1 system provides more consistent improvements over the MeanWT2012.

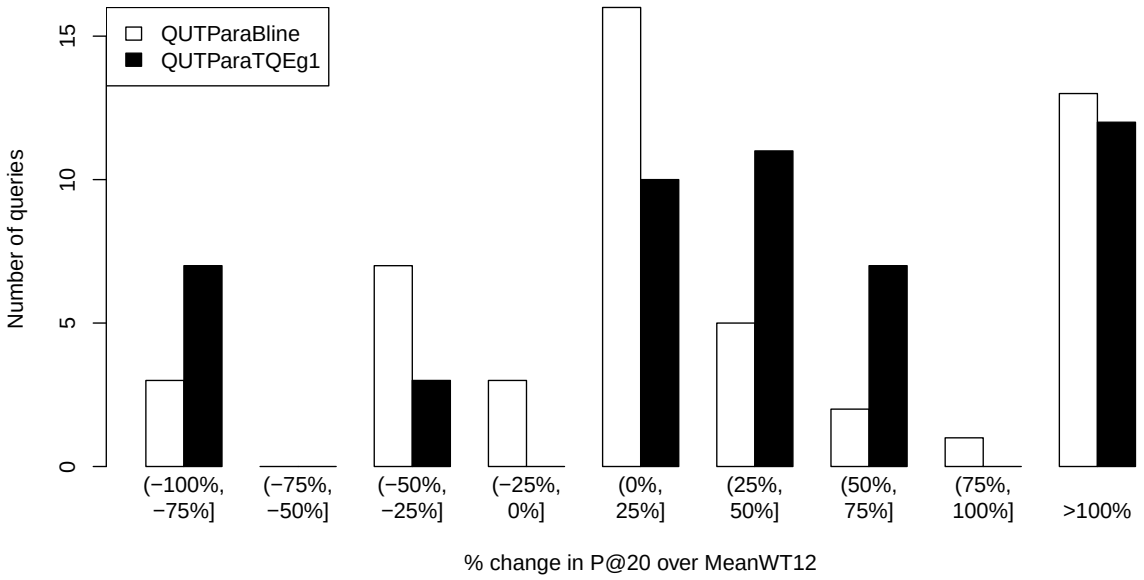

Figure 7.8: Robustness comparison of the QUTParaTQEg1 and QUTParaBline systems.

The increased variance of the TQE distribution shown in Figure 7.8 suggests that use of the same mix of syntagmatic and paradigmatic information on all test queries can have remarkably different impacts on retrieval effectiveness. This may indicate that for some queries insufficient vocabulary statistics exist to allow effective modelling of both syntagmatic and paradigmatic associations, as supported by the findings on the medical concept similarity judgement task (Section 4.4.2).

[11] P@20 was used as no MAP for MeanWT2012 was available, and Section 7.2.2 suggests that a comparison on ERR@20 or nDCG@20 is unlikely to be meaningful.

Evaluation Based on an Alternate Set of Relevance Judgements

When the results for the TREC Web track are initially released only the average, worst and best performance using ERR@20, nDCG@20 and P@20 across all submissions is provided; the individual submission results for other teams is not released. The average performance for all 48 submissions to the 2012 Web track is shown in Table 7.12, and a paired t-test confirms that both the QUTparaTQEg1 and QUTparaBline are both significantly better than these average metric scores at the 95% confidence interval, when using the NIST relevance judgements (i.e., for the official TREC scores).

However, after the TREC conference more details relating to each submissions rank on the ERR@20 and NDCG@20 measures are released. For 2012, these results included a ranking of submissions based on NIST relevance judgements as well as relevance judgements created by Microsoft Bing's proprietary Human Relevance System (HRS) [Craswell, 2012]. The rank of the QUTParaTQEg1 and QUTParaBline submissions for each of these relevance judgements are shown in Table 7.13, along with the performance on each measure.

	Measurement				Rank of system			
	NIST		HRS		NIST		HRS	
	NDCG	ERR	NDCG	ERR	NDCG	ERR	NDCG	ERR
QUTparaBline	0.167	0.290	0.334	0.324	15	9	5	3
QUTparaTQEg1	0.192	0.249	0.214	0.165	10	13	16	18

Table 7.13: Measurement and rank of the QUTParaTQEg1 and QUTParaBline submissions in the list of 48 submissions based on the NIST and Microsoft's HRS relavance judgements. Measures are @20.

It is not surprising to find the Google baseline submission moves up to rank 3 based on ERR@20 when the HRS relevance judgements are used, as these are based on user click-through data collected with Bing, Microsoft's search engine. There is a growing focus on better modelling of users in the information retrieval community [Allan et al., 2012]. This move is likely to see increasing use of user generated relevance judgements, such as user click-through logs, and way from the use of judgements created through system pooling, as is traditionally done in TREC, as judgements from pooling are often topic based.

The drop in ranking of the QUTParaTQEg1 when the HRS judgements are used raises questions as to whether the linguistic motivation of the TQE approach in using word meanings to augment query representations is able to model the user behavior captured in click-through logs. One suggestion may be that the rankings are based on the graded metric (ERR@20 and nDCG@20), and hence this drop in ranking may be more related to the ineffectiveness of the unigram language model to optimally rank the final results based on ERR, rather than the TQE approaches ability to effectively augment the query representation. Access to the MAP and P@20 scores produced for each submission based on the HRS rankings may provide better insight into any weakness the TQE approach may have in modelling user click-through behavior[12]. If the TQE approach is still achieving significant improvements in MAP and P@20 then significant improvements in ERR@20 and nDCG@20 over QUTparaBline are theoretically possible if the top 20 documents returned by QUTparaTQEg1 are more optimally ranked.

Diversity Results

As all ad hoc retrieval submissions were also evaluated on the diversity task, the diversity performance of our submission can be reported, and these results are shown in Table 7.14. These results show that both our submissions performed better than the average performance achieved by all 2012 TREC Web track participants.

It is worth noting, that no design choices were made in QUTparaTQEg1 aimed at improving the diversity of results. However, the significant improvements of TQE over both the average (MeanWT2012) and the Google baseline (QUTparaBline) on the P-IA@20 metric suggests that the combination of both paradigmatic and syntagmatic information may assist diversification of results. The performance of TQE (QUTparaTQEg1) on the graded metrics (i.e., α-nDCG@20 and ERR-IA@20) were not significantly different from the average (MeanWT2012) or the baseline (QUTparaBline).

7.2.3 Conclusion

In the QUT_Para submission to the TREC 2012 Web track, syntagmatic and paradigmatic associations between words were explicitly modelled within the TQE approach with the aim of improving retrieval effectiveness by more effectively augmenting query representations. The

[12]However, this data is not yet available, at time of publication

	α-nDCG@20	P-IA@20	ERR-IA@20
MeanWT2012	0.476	0.213	0.364
QUTparaBline	**0.527**^{m}	0.226	**0.419**^{m}
QUTparaTQEg1	0.498	**0.286**^{mb}	0.382

Table 7.14: Comparison of performance on TREC 2012 Web Track diversity task. The superscripts m, b and t indicate statistically significant differences (calculated using a paired t-test, $p < 0.05$) over the average performance of all TREC Web track participants (MeanWT2012), our baseline (QUTparaBline) and TQE approach (QUTparaTQEg1), respectively. The best results for each evaluation measure appear in boldface.

results indicate that the TQE approach can significantly improve retrieval effectiveness for binary metrics, such as MAP and P@20, and hence return more relevant documents in the top 20 when compared to the strong Google baseline submission.

A weakness with the QUTParaTQEg1 submission, which used the TQE approach, appears to be linked to the poor ability of the unigram language model to optimally rank the returned documents and hence provide significant improvements on graded measures, such as ERR@20 and nDCG@20. This highlights a potential area for future work, relating to the use of a learning to rank algorithm to assist with boosting these graded measures for the QUTParaTQEg1 system.

A final area for improvement, highlighted by this experiment, stems from the recurring observation that for very short queries TQE has difficulty modelling effective syntagmatic and paradigmatic information. The following section, relating to identifying the best mix of associations for a given query, may go some way toward addressing this issue, and potentially increasing the improvements in retrieval effectiveness of the TQE approach.

7.3 Oracle Analysis

The approach to evaluating the impact on retrieval effectiveness of combining syntagmatic and paradigmatic information within the query expansion process for the short and verbose experiments in this chapter, required all free parameters, except γ to be fixed. The mixing parameter, γ was then trained on MAP to find the best average retrieval effectiveness using a

3-fold cross validation approach. This training method was used as it is the accepted approach in the field of information retrieval. However, this approach effectively assumes there is no linguistic correlation between the queries and the value of γ. The same assumption is used by the QUTparaTQEg1 system in the Web track submission evaluated in Section 7.2.

Unlike probabilistic query expansion approaches that use statistical estimation techniques with little linguistic motivation, the TQE approach is grounded in linguistic theory and therefore some correlation between the best mix (γ) of syntagmatic and paradigmatic information and the query terms can be argued to exist. For example, paradigmatic associations are likely to benefit queries that suffer from vocabulary mismatch issues.

Evidence supporting this link between the query terms and the ability to choose whether syntagmatic or paradigmatic information would most effectively augment the query representation is provided by the findings of the medical concept similarity judgement experiment, reported in Section 4.4. These found that effective modelling of syntagmatic associations was dependent on the of within document co-occurrences of the medical concepts being compared, while the effective modelling of paradigmatic associations was dependent on the corpus frequencies of the concepts being considered. The same is proposed be true when modelling word associations for the query terms using the set of (pseudo) relevant documents.

A link between the vocabulary statistics that are present and the optimal mix of syntagmatic and paradigmatic associations would motivate the development of an adaptive TQE approach. Adaptive query expansion techniques have been successfully developed in the past using techniques from the field of machine learning [Lv and Zhai, 2009b]. To gauge how significant any improvements in the retrieval effectiveness of TQE may be if the optimal mix of syntagmatic and paradigmatic associations could be predicted, an oracle analysis can be performed. This will determine the theoretical upper limit in retrieval effectiveness of an adaptive TQE approach.

7.3.1 Short and Verbose Query Experiments

When this *oracle* approach is used to evaluate the TQE performance on the short and verbose ad hoc retrieval experiments presented earlier, significant improvements in retrieval effectiveness over the baseline on both MAP and P@20 are achieved, for a 99% confidence interval. These results are shown in Table 7.15.

It is worth noting, that this result is only the oracle performance of TQE approach for when

		short queries		verbose queries	
	Metric	noFB	TQE	noFB	TQE
WSJ	MAP	.2686	$.3184^{n}$ (18.5%)	.2121	$.3025^{n}$ (42.6%)
	P@20	.4074	$.4697^{n}$ (15.3%)	.3480	$.4593^{n}$ (32.0%)
AP	MAP	.1793	$.2473^{n}$ (37.9%)	.1511	$.2202^{n}$ (45.8%)
	P@20	.2300	$.3107^{n}$ (35.1%)	.2300	$.3248^{n}$ (41.2%)
ROB	MAP	.2500	$.2922^{n}$ (16.9.%)	.2491	$.3127^{n}$ (25.5%)
	P@20	.3558	$.4014^{n}$ (12.8%)	.3373	$.4131^{n}$ (22.4%)
G2	MAP	.2941	$.3304^{n}$ (13.22%)	.2466	$.2963^{n}$ (20.2%)
	P@20	.5050	$.5791^{n}$ (14.7%)	.4594	$.5419^{n}$ (18.0%)
CW	MAP	.0768	$.0881^{n}$ (14.8%)	.0530	$.06378^{n}$ (26.3%)
	P@20	.1872	$.2311^{n}$ (23.4%)	.1561	$.1959^{n}$ (25.5%)

Table 7.15: Oracle ad hoc retrieval results for TQE using short and verbose queries when compared to the unigram language model (noFB). ^{n} indicates statistically significant improvements ($p < 0.01$) over the baseline (noFB). Brackets indicate percent improvement over noFB.

γ is the only free parameter; and that the number of feedback documents, query expansion terms and mix of original query weights used within the pseudo relevance feedback setting were still fixed at 30, 30 and 0.5, respectively.

7.3.2 The 2012 TREC Web Track

	Metric	unigramLM	QUTParaBline	TQE
CW	MAP	.107	$.117^{u}$	$.178^{ub}$ (51.8%)
	P@20	.254	$.305^{u}$	$.434^{ub}$ (42.3%)

Table 7.16: Oracle ad hoc retrieval results for TQE on the 2012 TREC Web track compared to the unigram language model (unigramLM) and Google based QUTParaBline system. ^{u} and ^{b} indicate statistically significant improvements ($p < 0.01$) over the unigramLM and QUTParaBline approaches. Brackets indicate percent improvement over QUTParaBline.

When an oracle approach is applied to the TQE approach on the 2012 TREC Web track, the updated results for MAP and P@20 are shown in Table 7.16. These results indicate that if the optimal mix of syntagmatic and paradigmatic information in the TQE based approach could be predicted, then the percent improvement over the Google baseline would increase from 35% to 52% on MAP, and from 30% to 42% on P@20.

7.3.3 Conclusion

The oracle analysis performed on all three experiments relating to the evaluation of the TQE approach indicates that an adaptive TQE approach would provide significant improvements, within a 99% confidence interval, over the baseline models. This finding provides strong support for developing an adaptive TQE approach.

7.4 Summary

Information retrieval evaluation methodology does allow for the training of all free model parameters. However, when tuning all parameters it can be difficult to confidently evaluate the improvements in retrieval effectiveness provided by each of the parameters. This is why

there has been such a strong focus in this research on leaving the mix (γ) of syntagmatic and paradigmatic information as the only free parameter within the model. In this way we can precisely determine how much gain is achieved from using both types of linguistic associations.

By fixing all other free model parameters, except for the mix of syntagmatic and paradigmatic information, this approach, known as *tensor query expansion* (TQE), demonstrated significant improvements in ad hoc retrieval effectiveness over strong benchmark models for long queries across a wide range of data sets.

It appears that these long queries, also known as verbose queries, provide enough statistical information to ensure reliable modelling of paradigmatic associations, as opposed to shorter queries on which significant improvements in retrieval effectiveness were not always achieved. Effective modelling of paradigmatic associations and their use in the TQE approach appears to be responsible for producing these significant improvements.

Although restricting the number of free parameters in a model is a more rigorous approach, within the information retrieval community there is a strong interest in reporting significant improvements in retrieval effectiveness. This often leads to researchers training models with many free parameters. These types of evaluations are probably best suited to settings were the influence on any one parameter on effectiveness is not so important, as is often the case within the TREC forum. This is why the TQE approach was entered into the 2012 TREC Web track, and was allowed to have all free parameters trained. The results outlined in Section 7.2 demonstrated that the TQE approach could produce significant improvements in MAP and P@20 when compared to a very strong baseline. This baseline submission was formed using the Google retrieval service, and ranked third on ERR@20 in the 2012 TREC Web track when clickthrough relevance judgements were used, demonstrating that it was a very strong baseline system.

This chapter has rigorously evaluated the benefits of using the TE model to augment query representations within the information retrieval process. Information retrieval is one application where the TE model can provide significant improvements in task effectiveness. However, there may be many other problems were the TE model can be applied. A number of these will be discussed in the future work section of the following chapter. The concluding chapter will also summarise the major findings and contributions of this research along with directly addressing the original research questions posed in Chapter 1.

Part III

Concluding Remarks

Chapter 8

Conclusion and Future Work

This dissertation has proposed the *tensor encoding* (TE) model of word meaning, based on the structural linguistic view of meaning initially formulated by Swiss linguistic, Ferdinand de Saussure (1916). More generally, the TE model provides a formal framework for modelling the differences in meaning between concepts, be they single words, tuples or abstract concepts, based on their syntagmatic and paradigmatic associations with each other. Through rigorous evaluation on a broad range of semantic tasks, along with its application within the information retrieval process, the TE model has demonstrated the improvements in task effectiveness achieved by modelling word meaning in this way.

This chapter pulls together the threads of the story woven throughout this dissertation, and summarises them with respect to answering the research questions outlined in Chapter 1. The contributions of the research (Section 8.3) relate to several fields, including cognitive science and information retrieval. Finally, the future work (Section 8.4) highlights opportunities for further enhancing and applying the TE model.

8.1 Overview of the Research

The introduction to structural linguistic theory (Chapter 2) was used to motivate the development of a computational model of word meaning that combined a number of recent advances in *semantic space model* (SSM) technology. These advances included the use of structurally encoded, tensor representations and their storage in fixed dimension vectors. The way in which these advances could be combined into a formal model of word meaning, known as the *tensor encoding* (TE) model, was theoretically motivated from structural linguistics, and shown to be

computationally efficient when compared to existing corpus-based SSMs (Chapter 3).

The effectiveness of the TE model was evaluated on four semantic tasks, (i) synonym judgement, (ii) semantic distance, (iii) semantic categorization and (iv) medical concept similarity judgement (Chapter 4). The results demonstrate that the TE model's ability to flexibly mix information about syntagmatic and paradigmatic associations within the estimation process allows it to achieve superior task effectiveness when compared to a strong corpus-based SSMs. This is supported by the finding that optimal task effectiveness of the TE model on these tasks was achieved when both syntagmatic and paradigmatic information was used to underpin the similarity estimates. The robustness over a wide range of tasks was argued to allow the TE model to move beyond the *one model, one task* reputation associated with many existing corpus based SSMs.

The potential for SSMs to be applied on natural language tasks, such as information retrieval, motivated a review of information retrieval models and current approaches to query expansion (Chapter 5). The lack of paradigmatic information used within existing query expansion techniques and the dependence on word meanings within query formulation highlighted query expansion as a potentially fruitful application of the TE model. The TE model was formally applied within the relevance modelling framework, a popular query expansion approach with formal groundings, and was called the *tensor query expansion* (TQE) technique (Chapter 6).

Analysis of the TQE approach demonstrated that on longer queries, augmenting query representations with a mix of syntagmatic and paradigmatic information can provide significant improvements in ad hoc retrieval effectiveness across a wide range of data sets and compared to robust benchmark models (Chapter 7). The evaluation of TQE also showed that when all model parameters are tuned, significant improvements in retrieval effectiveness on web search can be achieved when compared to a strong baseline system built from an industry search engine (Chapter 7).

8.2 Addressing the Research Questions

The experimental results presented in this dissertation, and summarised in Section 8.1, provide evidence to allow responses to the original research questions (Chapter 1) to now be articulated.

1. **Can a corpus-based model of word meaning, formally combining syntagmatic and**

paradigmatic information, provide superior performance on semantic tasks when compared to current corpus-based SSMs? ANSWER: ***Yes***

The effectiveness of the TE model was evaluated on four semantic tasks, (i) synonym judgement, (ii) semantic distance, (iii) semantic categorization and (iv) medical concept similarity judgement (Chapter 4). The results demonstrate that the TE model can achieve superior effectiveness on these tasks when compared to a number of like and strong corpus-based SSMs. The finding that optimal task effectiveness was achieved when both syntagmatic and paradigmatic information is used, indicates that the TE model's formal framework is critical in achieving this result.

2. **Can a corpus-based model of word meaning, formally synthesising syntagmatic and paradigmatic information, be used to augment query representations and provide significant improvements in retrieval effectiveness over current information retrieval models?** ANSWER: ***Yes***

   The results of the independently assessed, TREC 2012 Web track demonstrate that the TQE approach achieves statistically significant improvements in retrieval effectiveness when compared to a strong baseline retrieval system (Section 7.2). Results on ad hoc retrieval tasks when only the mix of syntagmatic and paradigmatic information is varied within the TQE approach indicate that information about both forms of associations are required to achieve significant improvement in task effectiveness (Section 7.1.3). This finding, along with a similar result highlighted in the response to research question 1, strengthens the argument that the success of the TE model is due to its formal synthesis of syntagmatic and paradigmatic information.

8.3 Contributions

The contributions made by this research, include:

1. **The development of a new, unifying model of word meaning within which information about syntagmatic and paradigmatic associations between concepts can be formally synthesised.** Explicitly combining measures of syntagmatic and paradigmatic information within a formal framework was shown to provide superior effectiveness on the semantic tasks evaluated in this research (Chapter 4). This framework supports the use

of any measures of syntagmatic and paradigmatic association, whether they be corpus-based or from external linguistic resources. However, corpus-based approaches are preferred as they are generally cheaper and allow more context-sensitive modelling, which is consistent with theories of meaning proposed in structural linguistics (Section 2.2).

2. **The development of a novel compression technique that allows tensor representations to be efficiently stored in fixed dimension vectors.** The evolution of SSMs (Chapter 2) highlighted the forces behind the use of order-encoding approaches that store representations in fixed dimensions. Extending SSMs to use high-order tensor representations is intuitively costly, but has been argued to improve the effectiveness of SSMs across a broader range of applications. The development of the tensor memory compression (TMC) technique (Section 3.1.4) allows these advances in SSM technology to be combined efficiently within a single corpus-based SSM.

   It is worth noting that the TMC technique can be applied to other corpus-based SSMs that deal with sparse representations. The TMC technique enhances the ability of corpus-based SSMs to model temporal aspects of meaning (Section 2.2) and emulate the information compression that occurs within the human brain [Eliasmith et al., 2012].

3. **The development of a new measure of paradigmatic information that is less sensitive to the context window size.** The measure defined in Equation (3.30) has been shown to be an effective and robust measure of paradigmatic associations, which unlike most common measures of similarity will always primarily model paradigmatic associations, rather than syntagmatic associations, as the context window size is varied.

4. **The rigorous evaluation of a number of strong SSM benchmark models across a wide range of tasks.** Performing comparisons of models in the literature is often unreliable, due to the differing experimental setup. Therefore, this research implemented a number of benchmark corpus-based SSMs that use structural encoding, or have been shown to be the strongest benchmark to-date for corpus-based SSMs, and then undertook a robust evaluation of each with the TE model. This evaluation provides a side-by-side assessment of the merits of each model that could assist researchers in choosing an SSM for future research work.

5. **The evaluation of the benefits to retrieval effectiveness of modelling paradigmatic information within the information retrieval process.** Current information retrieval

models do not explicitly model syntagmatic or paradigmatic information within the retrieval process. The last decade has seen improvements in retrieval effectiveness through the use of syntagmatic information in augmenting query representations within the information retrieval process. The work presented in this dissertation demonstrates that further improvements in retrieval effectiveness can be made by explicitly modelling both syntagmatic and paradigmatic associations within the query expansion process.

This is of particular relevance to researchers in the field and industry as demonstrated by the results achieved at an international information retrieval conference, known as TREC, using TQE to significantly improve the retrieval effectiveness of a top industry search engine on a number of retrieval metrics (Section 7.2).

It is hoped that this work will also be seen as a significant contribution to the substantive dialogue between the fields of linguistics, cognitive science and information retrieval. Despite these new contributions, there remain many opportunities to gain an improved understanding of the potential benefits and drawbacks of the TE model and its applications. The following section outlines a number of areas for future work that may assist in developing this understanding.

8.4 Future Work

The first section below focuses primarily on enhancements that could be made to the TE model itself, and the second on future potential applications.

8.4.1 Enhancing the TE Model

Ideas for enhancing the TE model relate to improving effectiveness, efficiency and widening the range of semantic tasks on which the TE model could achieve robust effectiveness. These ideas could be considered to be extensions to the theoretical development of the model (i.e., extensions to Part I of this dissertation).

Developing an Adaptive TE Approach

An important finding observed when evaluating the effectiveness of using information about paradigmatic associations to (i) expand query representations on short queries (Section 7.1),

and (ii) judge the semantic similarity of medical concepts (Section 4.4.2) relates to the inability of the TE model to effectively model associations when insufficient vocabulary statistics exist. This link between vocabulary statistics and the effectiveness of paradigmatic information in assisting task effectiveness provides an opportunity to enhance the TE model to be more *adaptive*, i.e., choose the best mix of syntagmatic and paradigmatic information based on the vocabulary statistics for a given situation.

A second important finding from this research is that the most effective performance on any task is most often achieved when information about both syntagmatic and paradigmatic associations is used in combination, assuming sufficient vocabulary statistics exist to provide effective modelling of both associations. This link between superior task effectiveness and the inclusion of both types of association leads to an interesting question relating to the evolution of language:

> *If the use of syntagmatic and paradigmatic associations provides superior effectiveness on semantic tasks, would languages develop to ensure that both syntagmatic and paradigmatic associations exist between words?*

To answer this question, we need to know what property is required for both associations to exist. As effective paradigmatic associations are based on the occurrence patterns of close neighbouring words (Section 3.2.2), languages that ignore word order are unlikely to exhibit effective paradigmatic associations, because a measure of close proximity is unlikely to find strong patterns. Effective syntagmatic associations can exist between terms over a wide ranging context (Section 6.2.1), and hence languages ignoring word order may still model effective syntagmatic associations. This suggests a growing importance of word order within a language should develop.

An investigation into the evolution of word order traits within languages, found that most languages have a strong dependence on word order, the level of dependence varying across cultures [Dunn et al., 2011]. Our findings may provide a novel insight into why this is the case; i.e., as word order ensures languages contain paradigmatic information in addition to syntagmatic information, and this would provide improved effectiveness on tasks involving semantic judgements.

If the evolution of languages could have predicted the improved performance on semantic tasks when both syntagmatic and paradigmatic associations are modelled, then the next intuitive

question becomes:

> *How did the human brain determine the optimal mix of syntagmatic and paradigmatic information to use when processing language in real time?*

If this were known, then these mechanisms may provide a guide as to how best to develop an *adaptive* TE approach that could modify the mix of syntagmatic and paradigmatic information used when performing similarity judgements using the TE model. Based on the finding that effective modelling of syntagmatic and paradigmatic associations appears to be sensitive to characteristics of the TE vocabulary, it is likely that the mechanism behind an adaptive approach would be based on some vocabulary statistic of the words being compared. If humans estimate these word associations in a similar way, as researchers suggests they do [Jones and Mewhort, 2007], then they too are likely to use them to preference one association over another.

For example, results from the medical concept similarity judgements experiment (Section 4.4.2) showed that if either of the concepts in the pair being considered had very few occurrences in the training corpus, then syntagmatic associations should be preferred, and γ should be set to a value closer to zero (0) in the TE model. Alternatively, if there were no co-occurrence of the two concepts being considered within a single training document, then paradigmatic information should be used solely to estimate the similarity (i.e., $\gamma = 1$ in the TE model).

By using findings like these, query predictors or techniques from machine learning can be applied to predict the optimal mix of syntagmatic and paradigmatic information. Future work may hopefully determine whether an adaptive TE model can be created, and whether it would result in further improvements in task effectiveness.

Optimising the Measures used to Model Word Associations

A quick review of the measures chosen for modelling syntagmatic associations on the various tasks evaluated in this research, suggests they are often task specific. For example, on the task of ad hoc retrieval using pseudo relevance feedback, a measure based on the Dirichlet smoothed estimate of terms found in the set of pseudo relevant documents provides an efficient and effective estimate of syntagmatic associations. Meanwhile, on the task of medical concept similarity judgement the positive pointwise mutual information (PPMI) measure proved an

effective measure of syntagmatic associations. This result suggests that the most effective measures for modelling syntagmatic and paradigmatic associations within the TE framework may depend on a number of factors, including the nature of the task.

A second factor which impacts the choice of measure, and has been highlighted in past research, is the nature of the associations [Bullinaria and Levy, 2007, Xu and Croft, 1996]. Recall, syntagmatic associations often exist between words far apart in natural language [Xu and Croft, 1996], while paradigmatic associations are best modelled using co-occurrence information of close neighbours [Bullinaria and Levy, 2007]. Therefore, sourcing estimates off two semantic spaces, built using different context window sizes, may help improve task effectiveness.

A rigorous analysis of the most effective measures for modelling syntagmatic and paradigmatic associations on a wide range of tasks may allow a general rule for selecting measures to be created, and potentially lead to further improvements in task effectiveness.

Higher-order TE Models

Even though higher-order tensor representations have been shown to make SSMs more robust across a wider range of tasks [Baroni and Lenci, 2010], the order of the tensor representations that achieves optimal task effectiveness has not been rigorously evaluated. When evaluating the performance of the TE model on the semantic tasks outlined in Chapter 4, a second-order TE model demonstrated superior effectiveness over the benchmark models on all tasks. However, higher-order representations within the TE model could have been used, and the performance of these evaluated.

The same is true for the evaluation of the TQE technique (Chapter 7). Within query expansion, past efforts using information about n-tuples ($n > 1$) have failed to produce conclusive improvements [Metzler and Croft, 2007]. Therefore, a second-order TE model was used within TQE on all experiments, and hence only associations between individual words were modelled.

To understand the potential benefits of using higher-order representations within the TE model, evaluation of various order TE models needs to be undertaken, along with evaluation on a greater variety of semantic tasks, such as analogy making [Cohen et al., 2012, Turney, 2008]. As the formalism for the third-order variant of the TE model, along with details on implementing the storage vectors is provided (Chapter 3), this research forms a theoretical platform from which to undertake an evaluation comparing the performance of higher-order TE

models.

Given the likely increase in storage costs associated with higher-order representations, an investigation into optimal TMC algorithms (Section 3.1.4) may also be worthwhile. Research into effective policies for measuring informativeness have been developed for applications like entity detections [Rennie and Jaakkola, 2005] and index pruning [Carmel et al., 2001]. These may provide a sound starting point from which to develop an enhanced TMC algorithm.

The result of any future evaluation of higher-order TE models would hopefully identify task specific features which suggest the order of TE model that is most likely to provide optimal task effectiveness.

Incorporating External Linguistic Resources

The scope of this research intentionally avoided the use of external linguistic resources (i.e., POS taggers, parsers, manually built knowledge bases) so as to keep the TE model cheap and produce a generic model of meaning that could be applied across languages as well as outside linguistics. However, it is an accepted fact that a full account of human lexical knowledge will not likely come without the use of external linguistic knowledge. For this reason, the TE model presented in this dissertation is only a computational model of word meaning, and does not claim to be a psychological realistic model of human semantic processing.

Past research has shown that external linguistic resources can be successfully applied within corpus-based approaches [Baroni and Lenci, 2010, Mitchell and Lapata, 2010]. Lessons and design choices made within these research efforts could provide insight and guidance on how to enhance the TE model to incorporate external linguistic resources, extending it beyond the status of a "no frills" computational model.

8.4.2 Applications

The introduction of this dissertation provided a motivation for using SSM technology to underpin a formal model of word meaning that could be applied to natural language applications. This motivation was based on the reported likelihood that semantic technologies, like SSMs, may form a critical link in bridging the communication gap between humans and computers [Turney and Pantel, 2010]. The TE model of word meaning was then applied with success to the task of augmenting query representations within the information retrieval process (Part II).

As an extension to the second part of the research, the following section details several additional natural language applications to which the TE model could be applied, along with a discussion and example of applications outside of linguistics that may be well suited.

Query Expansion within Medical Information Retrieval Electronic medical information retrieval is increasingly important in areas such as evidence based medicine. This uses current best evidence to inform decisions about the care of individual patients. Identification of this best evidence often involves searching through large sources of medical literature.

Medical documents, such as medical literature and patient records make extensive use of domain specific language, including medical terms relating to diseases, drugs and treatments. To improve the consistency within a collection of medical documents they are often associated with representations based on ontological concepts by human expert coders. These concept based representations of documents were used in the medical concept similarity judgement task performed in Section 4.4.2.

The ability of the TE model to achieve a much greater correlation with human expert assessors when judging the similarity of medical concepts within medical documents compared to other corpus-based approaches (Section 4.4.2), along with its demonstrated effectiveness in expanding query representations within information retrieval (Chapter 7) provides strong motivation for applying the TE model to medical document retrieval.

The TQE formalism developed in Chapter 6 along with the concept based formalism of the TE model in Equation (3.24) could be used as a starting point to develop a concept based query expansion approach for medical retrieval. The updated form of the TQE approach in Equation (6.8) would become:

$$P_{G,\Gamma}(c_1|Q_c) = \frac{1}{Z_\Gamma}\left[(1-\gamma)s_{\text{syn}}(Q_c, c_1) + \gamma s_{\text{par}}(Q_c, c_1)\right], \tag{8.1}$$

where Q_c is the query represented as a sequence of medical concepts, $\gamma \in [0, 1]$ mixes the amount of syntagmatic and paradigmatic information used in the estimation, and Z_Γ is used to normalise the distribution.

Query Expansion for Geometric Retrieval Models The ad hoc retrieval experiments (Chapter 7) demonstrated the ability of the TE model to improve retrieval effectiveness by allowing access to both syntagmatic and paradigmatic information within the query expansion process.

This was achieved within a relevance modelling framework, primarily due to the formal grounding of the relevance model and the dominance of probabilistic models within information retrieval. However, the measures of syntagmatic and paradigmatic information within the TE model could be used to update query representations within a geometric based information retrieval model.

The Rocchio approach (Section 5.3.1) is one of the most popular and successful query expansion techniques for working with geometric representations of queries. Using the TE model to calculate the updated query weights, Equation (5.20) could be rewritten as:

$$q_j(1) = \alpha q_j(0) + \beta \frac{1}{|R|} S_{\text{TE}}(Q, q_j) - \gamma \frac{1}{|NR|} \sum_{D_i \in NR} d_{ij}, \tag{8.2}$$

where $S_{\text{TE}}(Q, q_j)$ is the TE model score for term q_j for the query Q and set of (pseudo) relevant documents R, and can be expressed as:

$$S_{\text{TE}}(Q, q_j) = (1 - \gamma) S_{\text{syn}}(Q, q_j) + \gamma S_{\text{para}}(Q, q_j), \tag{8.3}$$

where $S_{\text{syn}}(Q, q_j)$ and $S_{\text{para}}(Q, q_j)$ are some measures of syntagmatic and paradigmatic associations between Q and q_j.

Language Independent Information Retrieval With the increasing dependence on the Internet and the WWW by non-English speakers there is a need for information retrieval systems to be able to search for pages in any language. Many natural language processing approaches available to information retrieval systems come with the increased cost of requiring hand crafted linguistic resources, such as dictionaries and thesauri. However, as corpus-based SSMs, including the TE model, build their representations of words independently from the language of the underlying training documents, they provide a cheap tool for building language independent information retrieval systems.

It is worth noting that syntagmatic and paradigmatic associations between linguistic concepts are reported to exist to some degree within most languages; including Chinese, Japanese, Russian and Latin based languages used throughout Europe and Africa due to colonisation. Therefore, the differential view of meaning, that underpins the TE model, has broad application across natural languages.

Another increasingly important function of information retrieval systems is the ability to allow users to enter queries in a language that is different from that used within the documents

being searched, or even across a set of document written in multiple languages. SSMs are highly suited to this task of cross-language retrieval and have been successfully used in the past [Littman et al., 1998, Shakery and Zhai, 2013]. Given the TE model's demonstrated effectiveness when augmenting query representations in the information retrieval process (Chapter 7) and its superior effectiveness to several corpus-based SSMs (Chapter 4) it may be well suited for cross-language retrieval.

Semantic Tagging Knowledge management increasingly relies on the need to process semantic information. It is envisaged that the increased maturity and adoption of semantic technologies, like SSMs, will enable two-thirds of (manual) information management related tasks to be automated [Friedman et al., 2009].

One particular such need stems from the development of the "semantic web" [Shadbolt et al., 2005], which relies on a categorization process, known as *semantic tagging*, to associate meaningful information with web pages. This information can then be used to allow more effective search, data integration and the like [Raden, 2005].

Given the demonstrated effectiveness of the TE model on semantic categorization experiments within this research and the reliance of the semantic tagging process on this type of categorization, applying the TE model in this context may have potential benefits.

Literature-based Discovery Literature-based discovery refers to the use of literature to discover new relationships between existing knowledge or concepts. The methods underpinning literature-based discovery often rely on indirect connections between concepts. For example, the idea of linking concept A and C together because they each have a relationship with concept B, but not with each other is referred to as *Swanson linking*, and is based on the use of word co-occurrence information [Stegmann and Grohmann, 2003]. This approach has been used to underpin successful applications for investigating novel links between concepts within the medical domain [Smalheiser and Swanson, 1998].

The types of *second-order* relationships[1] used by Swanson linking are akin to those on which paradigmatic associations are based. Given the TE model's ability to explicitly combine information about both syntagmatic and paradigmatic associations and the success of SSMs on this task in the past [Cohen et al., 2010], the TE model may prove to be an effective approach

[1] Not to be confused with second-order representation referred to in this research.

to literature-based discovery.

Summary This list of potential applications for the TE model is restricted to natural language tasks and is by no means comprehensive. Briefly, others include recommender systems, spam detection and other data mining applications. However, the types of relationships modelled within the TE model have been shown to exist in domains outside of linguistics.

Applications Outside of Linguistics

The syntagmatic and paradigmatic associations that underpin the differential view of meaning, articulated by Ferdinand de Saussure, are based on the co-occurrence patterns of linguistic concepts (i.e., words, noun phrases, and others) (Section 2.2). However, within semiotics, which is the study of signs, linguistic units are not the only type of sign. In fact a sign is any physical form that can be imagined or made externally to stand for an object, event, etc.

The differential view of meaning, based on the existence of syntagmatic and paradigmatic associations between signs, can be argued to exist within domains found outside of linguistics. Being able to cheaply extract meaning from the interaction of concepts using this differential view of meaning underpinning the TE model is likely to be an appealing feature, especially in environments where the amount of unstructured data available makes manual analysis impractical.

For instance, consider the interactions between social media users. A synagmatic association could be argued to exist between user A and user B, if A posts on B's social media site. A paradigmatic association could be argued to exist between user A and user C, if C also posts on B's social media site. It could be reasonably deduced that user A and user C both know of user B, however, A and C may be unaware of each other's existence if they have not previously posted on each other's site.

The growth of social media sites and the explosion in user created content provides a rich source of information that can be used by many interested parties, including companies looking to improve their e-commerce sales or law enforcement agencies working on anti-terrorism strategies. Social network and mining applications have found that syntagmtatic and paradigmatic relationships between social media users are a valuable source of information. These relationships have been modelled using geometric models in the past, with emphasis

being placed on the ability to access implicit (paradigmatic) relationships between users in addition to direct (syntagmatic) [Yang et al., 2012]. The TE model fits this scenario well, as it provides a cheap and formal framework within which to model and combine information about syntagmatic and paradigmatic relationships between social media users.

This example, using social media users, combined with the theoretical development of a general TE model (Section 3.2.1) demonstrate the TE model's potential to become a *formal model of meaning*[2] that can estimate the similarity in meaning of abstract concepts that are non-uniformly distributed within an environment.

8.5 Final Remarks

Two of the most significant virtues of the TE model include: (i) the model's mathematical formalism, and (ii) its theoretical grounding in structural linguistics. These virtues enable the model to be extended and applied to a wide range of contexts, within which a differential view of meaning can be induced solely from the interactions of abstract concepts in the environment. Combined with the computational efficiency and task effectiveness demonstrated in this research, the TE model can be argued to provide a cheap, flexible framework for extracting meaning from the environment.

[2]Note, not just a model of *word* meaning

Appendix A

A Cosine Measure for the Second-order TE model

This appendix outlines the proof of an efficient cosine metric that can be used to measure the similarity of two representations within the second-order TE model.

A.1 Cosine Measure for Matrices

Consider two n-by-n matrices, A and B:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \\ a_{n1} & \dots & a_{nn} \end{pmatrix}, B = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \\ b_{n1} & \dots & b_{nn} \end{pmatrix},$$

that can also be expressed algebraically in terms of outer products as:

$$A = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} e_i e_j^T \quad , \quad B = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} e_i e_j^T, \tag{A.1}$$

where e_i and e_j are unit vectors and e_j^T indicates the transpose of the unit vector e_j. In the real inner-product space $\mathbb{R}^{n\times n}$, the ratio measure of the angle between the two nonzero matrices $A, B \in \mathbb{R}^{n\times n}$ is defined to be the number $\theta \in [0, \pi]$ such that

$$\cos = \frac{< A, B >}{||A||_F ||B||_F}, \tag{A.2}$$

where $< A, B >= tr(A^T B)$ is an inner product defined on $\mathbb{R}^{n\times n}$ and for example, $||A||_F = \sqrt{< A, A >} = \sqrt{tr(A^T A)}$, with $tr(.)$ representing the trace of the matrix. Given that:

$$tr(A^T B) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} b_{ij}, \tag{A.3}$$

and

$$tr(A^T A) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2, \tag{A.4}$$

$$tr(B^T B) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij}^2, \tag{A.5}$$

then Equation (A.2) can be re-written as:

$$\cos = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2} \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij}^2}}. \tag{A.6}$$

Note that Equation (A.6) has polynomial time complexity with respect to n, the number of rows and columns in the matrices.

A.2 Cosine Measure for Memory Matrices

Recall that after building the vocabulary, using the rank 2 Tensor Model's binding process, each term will have a memory matrix of the following form:

$$M_w = \begin{pmatrix} 0, & \ldots, 0, & f_{1w}, & 0, & \ldots, 0 \\ & & \ldots & & \\ 0, & \ldots, 0, & f_{(w-1)w}, & 0, & \ldots, 0 \\ f_{w1}, \ldots, f_{w(w-1)}, & & f_{ww}, & f_{w(w+1)}, \ldots, f_{wn} & \\ 0, & \ldots, 0, & f_{(w+1)w}, & 0, & \ldots, 0 \\ & & \ldots & & \\ 0, & \ldots, 0, & f_{nw}, & 0, & \ldots, 0 \end{pmatrix},$$

where f_{iw} (found in column w, row i of the memory matrix) is the co-occurrence frequency of term i preceding term w within the sliding context window over the text, f_{wi} (found in row w, column i of the matrix) is the co-occurrence frequency of term i succeeding term w within the sliding context window over the text, w and i are terms in the vocabulary V, and the vocabulary consists of n terms, such that $|V| = n$.

This memory matrix for term w, can be represented using dyadic notation:

$$M_w = \sum_{i \in V} f_{iw} e_i e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wj} e_w e_j^T, \tag{A.7}$$

where the first sum term represents the elements in column w of the matrix, the second sum term represents the values in row w of the matrix (excluding f_{ww} as it is already in the first sum term), and e_i, e_j and e_w are unit vectors.

To provide a proof of a cosine measure that is able to be used in applications of the TE model where a set of ordered terms are involved, as in the case of query expansion (Chapter 6), Q is introduced. Q represents an ordered set (known as a sequence) of vocabulary terms $Q = (q_1, \ldots, q_p)$ and within the implementation of the second-order TE model in this research (Chapter 3) is formed by summing the memory matrices of the terms in Q, such that:

$$M_Q = M_{q_1} + \ldots + M_{q_p}.$$

In dyadic notation, this is written as:

$$M_Q = \sum_{a \in V} f_{aq_1} e_a e_{q_1}^T + \sum_{bin\{V | b \neq q_1\}} f_{q_1 b} e_{q_1} e_b^T + \ldots + \sum_{g \in V} f_{gq_p} e_g e_{q_p}^T + \sum_{h \in \{V | h \neq q_p\}} f_{q_p h} e_{q_p} e_h^T. \tag{A.8}$$

To calculate the cosine measure for the memory matrix of term w (i.e., M_w) and the memory matrix of the terms in Q (i.e., M_Q), Equation (A.2) can be re-written as:

$$cos = \frac{tr(M_Q^T . M_w)}{\sqrt{tr(M_Q^T . M_Q)} . \sqrt{tr(M_w^T . M_w)}}. \tag{A.9}$$

Firstly, the term $M_Q^T . M_w$ inside the trace of the numerator, can be expressed algebraically as:

$$M_Q^T . M_w = \left[\sum_{a \in V} f_{aq_1} e_a e_{q_1}^T + \sum_{b \in \{V | b \neq q_1\}} f_{q_1 b} e_{q_1} e_b^T + \ldots + \sum_{g \in V} f_{gq_p} e_g e_{q_p}^T + \sum_{h \in \{V | h \neq q_p\}} f_{q_p h} e_{q_p} e_h^T \right]^T \times$$

$$\left[\sum_{i \in V} f_{iw} e_i e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wj} e_w e_j^T \right].$$

Transposing matrix M_Q:

$$= \left[\sum_{a \in V} f_{aq_1} e_{q_1} e_a^T + \sum_{b \in \{V | b \neq q_1\}} f_{q_1 b} e_b e_{q_1}^T + \ldots + \sum_{g \in V} f_{gq_p} e_{q_p} e_g^T + \sum_{h \in \{V | h \neq q_p\}} f_{q_p h} e_h e_{q_p}^T \right] \times$$

$$\left[\sum_{i \in V} f_{iw} e_i e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wj} e_w e_j^T \right].$$

Multiplying M_w through the terms of M_Q:

$$= \sum_{a \in V} f_{aq_1} e_{q_1} e_a^T \times \left[\sum_{i \in V} f_{iw} e_i e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wj} e_w e_j^T \right] +$$

$$\sum_{b \in \{V | b \neq q_1\}} f_{q_1 b} e_b e_{q_1}^T \times \left[\sum_{i \in V} f_{iw} e_i e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wj} e_w e_j^T \right] + \ldots$$

$$+ \sum_{g \in V} f_{gq_p} e_{q_p} e_g^T \times \left[\sum_{i \in V} f_{iw} e_i e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wj} e_w e_j^T \right] +$$

$$\sum_{h \in \{V | h \neq q_p\}} f_{q_p h} e_h e_{q_p}^T \times \left[\sum_{i \in V} f_{iw} e_i e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wj} e_w e_j^T \right].$$

Expanding:

$$= \sum_{a \in V} \sum_{i \in V} f_{aq_1} f_{iw} e_{q_1} (e_a^T e_i) e_w^T + \sum_{a \in V} \sum_{j \in \{V | j \neq w\}} f_{aq_1} f_{wj} e_{q_1} (e_a^T e_w) e_j^T +$$

$$\sum_{b \in \{V | b \neq q_1\}} \sum_{i \in V} f_{q_1 b} f_{iw} e_b (e_{q_1}^T e_i) e_w^T + \sum_{b \in \{V | b \neq q_1\}} \sum_{j \in \{V | j \neq w\}} f_{q_1 b} f_{kj} e_b (e_{q_1}^T e_w) e_j^T + \ldots +$$

$$\sum_{g \in V} \sum_{i \in V} f_{gq_p} f_{iw} e_{q_p} (e_g^T e_i) e_w^T + \sum_{g \in V} \sum_{j \in \{V | j \neq w\}} f_{gq_p} f_{wj} e_{q_p} (e_g^T e_w) e_j^T +$$

$$\sum_{h \in \{V | b \neq q_1\}} \sum_{i \in V} f_{q_p h} f_{iw} e_h (e_{q_p}^T e_i) e_w^T + \sum_{h \in \{V | h \neq q_p\}} \sum_{j \in \{V | j \neq w\}} f_{q_p h} f_{wj} e_h (e_{q_p}^T e_w) e_j^T.$$

Using the Kronecker delta, defined as:

$$e_g^T e_f = \delta_{gf} = \begin{cases} 0 & \text{if } g \neq f \\ 1 & \text{if } g = f \end{cases}$$

the following variables can be rationalised:

$$(e_a^T e_i) \rightarrow \quad a = i$$

$$(e_a^T e_w) \rightarrow \quad a = w$$

$$(e_{q_1}^T e_i) \rightarrow \quad q_1 = i$$

$$(e_{q_1}^T e_w) \rightarrow \quad q_1 = w$$

$$(e_g^T e_i) \rightarrow \quad g = i$$

$$(e_g^T e_w) \rightarrow \quad g = w$$

$$(e_{q_p}^T e_i) \rightarrow \quad q_p = i$$

$$(e_{q_p}^T e_w) \rightarrow \quad q_p = w$$

hence, the trace term, $M_Q^T.M_w$ becomes:

$$= \sum_{a \in V} f_{aq_1} f_{aw} e_{q_1} e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wq_1} f_{wj} e_{q_1} e_j^T + \sum_{b \in \{V | b \neq q_1\}} f_{q_1 b} f_{q_1 w} e_b e_w^T +$$

$$\sum_{b \in \{V | b \neq w, w = q_1\}} \sum_{j \in \{V | j \neq w, w = q_1\}} f_{wb} f_{wj} e_b e_j^T + \ldots + \sum_{g \in V} f_{gq_p} f_{gw} e_{q_p} e_w^T + \sum_{j \in \{V | j \neq w\}} f_{wq_p} f_{wj} e_{q_p} e_j^T +$$

$$\sum_{h \in \{V | b \neq q_1\}} f_{q_p h} f_{q_p w} e_h e_w^T + \sum_{h \in \{V | h \neq w, w = q_p\}} \sum_{j \in \{V | j \neq w, w = q_p\}} f_{wh} f_{wj} e_h e_j^T .$$

This is the general result of the trace term formed by M_Q and M_w. Taking the trace of this result requires summing the diagonal elements. These occur when the unit vectors of each term are equal (i.e., $q_1 = w$ in the first term, $q_1 = j$ in the second term, etc). So the trace of $M_Q^T.M_w$ becomes:

$$tr(M_Q, M_w) = \sum_{a \in \{V | w = q_1\}} f_{aw}^2 + (f_{wq_1}^2)_{w \neq q_1} + (f_{q_1 w}^2)_{w \neq q_1} + \sum_{j \in \{V | j \neq w, w = q_1\}} f_{wj}^2 + \ldots$$

$$+ \sum_{g \in \{V | w = q_p\}} f_{gw}^2 + (f_{wq_p}^2)_{w \neq q_p} + (f_{q_p w}^2)_{w \neq q_p} + \sum_{j \in \{V | j \neq w, w = q_p\}} f_{wj}^2 .$$

A definite pattern has formed, and grouping like terms for each term in Q gives:

$$tr(M_Q, M_w) = \sum_{a \in \{V | w = q_1\}} f_{aw}^2 + \sum_{j \in \{V | j \neq w, w = q_1\}} f_{wj}^2 + (f_{wq_1}^2 + f_{q_1 w}^2)_{w \neq q_1} + \ldots$$

$$\sum_{g \in \{V | w = q_p\}} f_{gw}^2 + \sum_{j \in \{V | j \neq w, w = q_p\}} f_{wj}^2 + (f_{wq_p}^2 + f_{q_p w}^2)_{w \neq q_p} .$$

Summing over all terms in Q, the resulting general expression for the trace of M_Q and M_w becomes:

$$tr(M_Q, M_w) = \sum_{i \in \{Q|i=w\}} (\sum_{j \in V} f_{ji}^2 + \sum_{j \in \{V|j \neq i\}} f_{ij}^2) + \sum_{i \in \{Q|w \neq i\}} (f_{wi}^2 + f_{iw}^2), \tag{A.10}$$

which reduces to:

$$tr(M_Q, M_w) = \sum_{j \in \{V|w \in Q\}} f_{ji}^2 + \sum_{j \in \{V|j \neq i, w \in Q\}} f_{ij}^2 + \sum_{i \in \{Q|w \neq i\}} (f_{wi}^2 + f_{iw}^2). \tag{A.11}$$

A similar process can be used to simplify the trace expressions in the denominator of Equation (A.9). The resulting trace expressions become:

$$tr(M_Q, M_Q) = \sum_{i \in Q} \left(\sum_{j \in V} f_{ji}^2 + \sum_{j \in \{V|j \neq i\}} f_{ij}^2 \right), \tag{A.12}$$

and

$$tr(M_w, M_w) = \sum_{j \in V} f_{jw}^2 + \sum_{j \in \{V|j \neq w\}} f_{wj}^2. \tag{A.13}$$

By substituting Equation (A.11), Equation (A.12), and Equation (A.13) into Equation (A.9), the cosine measure for memory matrices M_Q and M_w becomes:

$$\cos = \frac{< M_Q, M_w >}{||M_Q||_F ||M_w||_F} = \frac{\sum_{j \in \{V|w \in Q\}} f_{jw}^2 + \sum_{j \in \{V|j \neq w, w \in Q\}} f_{wj}^2 + \sum_{i \in \{Q|i \neq w\}} (f_{iw}^2 + f_{wi}^2)}{\sqrt{\sum_{i \in Q} \left[\sum_{j \in V} f_{ji}^2 + \sum_{j \in \{V|j \neq i\}} f_{ij}^2 \right]} \sqrt{\sum_{j \in V} f_{jw}^2 + \sum_{j \in \{V|j \neq w\}} f_{wj}^2}}. \tag{A.14}$$

When compared to Equation (A.6), it can be seen that the time complexity of the cosine metric for matrices has been reduced from polynomial order, i.e., $O(n^2)$ to linear order, i.e., $O(n|Q|)$, where $n = |V|$, the number of terms in the vocabulary, and $|Q|$ is the number of elements in the sequence Q.

A.2.1 An Efficiency Improvement

At an implementation level, the time complexity of Equation (A.14) can be reduced further, by taking advantage of any redundancy that may exist in Q. In the instance where the same term(s) may appear multiple times in Q, applying a multiplying factor to the original dyadic form of M_Q will avoid having to loop through all vocabulary terms unnecessarily. In this case, the original representation of Q=$(q_1, \ldots, q_p)$ could have been expressed as:

$$M_Q = s_{q_1}.M_{q_1} + \ldots + s_{q_m}.M_{q_m},$$

where Q has m unique terms $Q_m = \{q_1, \ldots, q_m\}$, and s_{q_i} represents the number of times q_i appears in the original squence, Q.

The dyadic notation for M_Q would have been expressed as:

$$M_Q = \sum_{a \in V} s_{q_1}.f_{aq_1} e_a e_{q_1}^T + \sum_{b \in \{V | b \neq q_1\}} s_{q_1}.f_{q_1 b} e_{q_1} e_b^T + \ldots + \sum_{g \in V} s_{q_m}.f_{g q_m} e_g e_{q_m}^T + \sum_{h \in \{V | h \neq q_m\}} s_{q_m}.f_{q_p h} e_{q_m} e_h^T. \tag{A.15}$$

Using this revision, Equation (A.11) becomes:

$$tr(M_Q, M_w) = \sum_{j \in \{V | w \in Q_m\}} s_i^2.f_{ji}^2 + \sum_{j \in \{V | j \neq i, w \in Q_m\}} s_i^2.f_{ij}^2 + \sum_{i \in \{Q_m | w \neq i\}} (s_i^2.f_{wi}^2 + s_i^2.f_{iw}^2). \tag{A.16}$$

A similar process can be followed for the magnitude of the trace in Equation (A.12):

$$tr(M_Q, M_Q) = \sum_{i \in Q_m} \left(\sum_{j \in V} s_i^2.f_{ji}^2 + \sum_{j \in \{V | j \neq i\}} s_i^2.f_{ij}^2 \right), \tag{A.17}$$

Substituting Equation (A.16) and Equation (A.17) into Equation (A.9) gives:

$$\cos\theta = \frac{\sum_{j \in \{V | w \in Q_m\}} s_i^2.f_{jw}^2 + \sum_{j \in \{V | j \neq w, w \in Q_m\}} s_i^2.f_{wj}^2 + \sum_{i \in \{Q_m | i \neq w\}} (s_i^2.f_{iw}^2 + s_i^2.f_{wi}^2)}{\sqrt{\sum_{i \in Q_m} \left[\sum_{j \in V} s_i^2.f_{ji}^2 + \sum_{j \in \{V | j \neq i\}} s_i^2.f_{ij}^2 \right]} \sqrt{\sum_{j \in V} f_{jw}^2 + \sum_{j \in \{V | j \neq w\}} f_{wj}^2}}. \tag{A.18}$$

A.2.2 Cosine of Two Memory Matrices

In the case where Q is a single term, i.e., $Q = (q)$, Equation (A.14) becomes:

$$\cos\theta = \frac{\sum_{j \in \{V | j \neq w, w = q\}} (f_{jq}^2 + f_{qj}^2) + (f_{qw}^2 + f_{wq}^2)}{\sqrt{\sum_{j \in V} f_{jq}^2 + \sum_{jin \{V | j \neq q\}} f_{qj}^2} \sqrt{\sum_{j \in V} f_{jw}^2 + \sum_{j \in \{V | j \neq w\}} f_{wj}^2}}. \tag{A.19}$$

Appendix B

Stoplist for Evaluation of the TE model

Table B.1 lists the stop words that made up the stoplist used on the synonym judgement part of TOEFL, semantic distance and semantic categorization tasks performed in Chapter 4.

a	can	go	might	respectively	thru
about	cannot	gone	more	s	thus
above	cant	got	moreover	second	to
accordingly	cause	h	most	secondly	too
across	causes	had	mostly	seem	twice
after	certain	has	mr	seemed	two
afterwards	changes	have	much	seeming	u
again	co	having	must	seems	unless
against	come	he	n	selves	until
all	contains	hence	name	sent	unto
allows	corresponding	her	namely	seven	up
almost	could	here	near	several	us

Continued on next page

along	currently	hereafter	necessary	shall	used
already	d	hereby	neither	should	uses
also	day	herein	never	since	using
although	did	hereupon	nevertheless	six	usually
always	do	hither	nine	so	v
am	does	how	no	some	various
among	doing	howbeit	nobody	somebody	very
amongst	done	however	none	somehow	via
an	e	i	noone	someone	viz
and	each	ie	nor	something	vs
another	eg	if	not	sometime	w
any	eight	ignored	nothing	sometimes	was
anybody	either	immediate	novel	somewhat	way
anyhow	else	in	nowhere	somewhere	we
anyone	elsewhere	inasmuch	o	specified	were
anything	et	inc	of	specify	what
anywhere	etc	indeed	off	specifying	whatever
apart	even	indicated	oh	state	when
appropriate	ever	indicates	on	sub	whence
are	every	inner	once	such	whenever

Continued on next page

around	everybody	insofar	one	sup	whereafter
as	everyone	instead	ones	t	whereas
aside	everything	into	only	taken	whereby
associated	everywhere	inward	onto	than	wherein
at	ex	is	or	that	whereupon
available	except	it	other	the	wherever
away	f	its	others	then	whether
awfully	far	itself	otherwise	thence	which
b	few	j	ought	there	whither
be	fifth	just	overall	thereafter	whoever
became	first	k	p	thereby	whole
because	five	keep	per	therefore	whom
become	for	kept	perhaps	therein	whose
becomes	forth	l	placed	thereupon	will
becoming	four	last	please	these	with
been	from	latter	plus	third	within
beforehand	further	latterly	q	this	without
besides	furthermore	lest	que	thorough	would
both	g	ltd	quite	thoroughly	x
but	get	m	r	those	y

Continued on next page

by	gets	many	rather	though	yet
c	given	may	really	three	z
came	gives	meanwhile	relatively	throughout	zero

Table B.1: Stoplist used for TOEFL, semantic distance and semantic categorization experiments in Chapter 4.

Appendix C

Data Sets for Medical Concepts

This appendix outlines the *Ped* and *Cav* data sets used for the similarity judgement of medical concepts experiment carried out in Chapter 4, Section 4.4.

C.1 Pedersen Data Set

Table C.1 outlines the concepts used to make up the Ped data set, based on those outlined in Pedersen et al. [Pedersen et al., 2007]. This includes the similarity judgements provided by 3 physicians and 9 clinical terminologists (coders), with the most similar terms scoring a 4 and least similar a 1.

Concept 1	**Concept 2**	**Physician**	**Coders**
C0035078 (Renal failure)	C0035078 (Kidney failure)	4	4
C0018787 (Heart)	C0027051 (Myocardium Infarction)	3.3	3
C0038454 (Stroke)	C0021308 (Infact)	3	2.8
C0156543 (Abortion)	C0000786 (Miscarriage)	3	3.3
C0011253 (Delusion)	C0036341 (Schizophrenia)	3	2.2

Continued on next page

Concept 1	**Concept 2**	**Physician**	**Coders**
C0018802 (Congestive heart failure)	C0034063 (Pulmonary edema)	3	1.4
C0027627 (Metastasis)	C0001418 (Adenocarcinoma)	2.7	1.8
C0175895 (Calcification)	C0009814 (Stenosis)	2.7	2
C0011991 (Diarrhea)	C0344375 (Stomach cramps)	2.3	1.3
C0026269 (Mitral stenosis)	C0004238 (Atrial fibrillation)	2.3	1.3
C0003873 (Rheumatoid arthritis)	C0409974 (Lupus)	2	1.1
C0006118 (Brain tumor)	C0151699 (Intracranial hemorrhage)	2	1.3
C0007286 (Carpel tunnel syndrome)	C0029408 (Osteoarthritis)	2	1.1
C0011849 (Diabetes mellitus)	C0020538 (Hypertension)	2	1
C0702166 (Acne)	C0039142 (Syringe)	2	1
C0003232 (Antibiotic)	C0020517 (Allergy)	1.7	1.2
C0010137 (Cortisone)	C0086511 (Total knee replacement)	1.7	1
C0034065 (Pulmonary embolus)	C0027051 (Myocardial infarction)	1.7	1.2
C0034069 (Pulmonary fibrosis)	C0242379 (Lung cancer)	1.7	1.4
C0206698 (Cholangiocarcinoma)	C0009378 (Colonoscopy)	1.3	1
C0026769 (Multiple sclerosis)	C0033975 (Psychosis)	1	1
C0003615 (Appendicitis)	C0029456 (Osteoporosis)	1	1

Continued on next page

Concept 1	Concept 2	Physician	Coders
C0034887 (Rectal polyp)	C0003483 (Aorta)	1	1
C0043352 (Xerostomia)	C0023891 (Alcoholic cirrhosis)	1	1
C0030920 (Peptic ulcer disease)	C0027092 (Myopia)	1	1
C0011581 (Depression)	C0007642 (Cellulites)	1	1
C0042345 (Varicose vein)	C0224701 (Entire knee meniscus)	1	1
C0020473 (Hyperlidpidemia)	C0027627 (Metastasis)	1	1

Table C.1: Medical Concept Pairs provided by Pedersen et al. [Pedersen et al., 2007].

C.2 Caviedes and Cimino Data Set

Table C.2 outlines the concepts used to make up the **Cav** data set, based on those outlined in Caviedes and Cimino [Caviedes and Cimino, 2004]. This includes the similarity judgements provided by 3 physicians, with the most similar terms scoring a 10 and least similar a 1.

Concept 1	Concept 2	Physician
C0002962 (Angina Pectoris)	C0010068 (Coronary Disease)	8.47
C0003811 (Arrhythmia)	C0018799 (Heart Diseases)	8.17
C0007192 (Cardiomyopathy, Alcoholic)	C0018799 (Heart Diseases)	8.17
C0010068 (Coronary Disease)	C0018799 (Heart Diseases)	8.17
C0018799 (Heart Diseases)	C0018802 (Heart Failure, Congestive)	8.17

Continued on next page

Concept 1	**Concept 2**	**Physician**
C0002962 (Angina Pectoris)	C0018799 (Heart Diseases)	7.17
C0018834 (Heartburn)	C0000737 (Abdominal pain)	7.17
C0010068 (Coronary Disease)	C0018802 (Heart Failure, Congestive)	6.97
C0007192 (Cardiomyopathy, Alcoholic)	C0018802 (Heart Failure, Congestive)	6.84
C0003811 (Arrhythmia)	C0018802 (Heart Failure, Congestive)	6.67
C0003811 (Arrhythmia)	C0007192 (Cardiomyopathy, Alcoholic)	6.5
C0003811 (Arrhythmia)	C0010068 (Coronary Disease)	6.3
C0002962 (Angina Pectoris)	C0003811 (Arrhythmia)	6
C0002962 (Angina Pectoris)	C0018802 (Heart Failure, Congestive)	6
C0003811 (Arrhythmia)	C0020621 (Hypokalemia)	5.84
C0002962 (Angina Pectoris)	C0007192 (Cardiomyopathy, Alcoholic)	5.67
C0007192 (Cardiomyopathy, Alcoholic)	C0010068 (Coronary Disease)	5.67
C0018802 (Heart Failure, Congestive)	C0020621 (Hypokalemia)	4.67
C0002962 (Angina Pectoris)	C0018834 (Heartburn)	4.34
C0018799 (Heart Diseases)	C0018834 (Heartburn)	3.67
C0000737 (Abdominal pain)	C0035238 (Respiratory System Abnormalities)	3.34
C0007192 (Cardiomyopathy, Alcoholic)	C0018834 (Heartburn)	3.34
C0010068 (Coronary Disease)	C0018834 (Heartburn)	3.34
C0018799 (Heart Diseases)	C0020621 (Hypokalemia)	3.34

Continued on next page

Concept 1	**Concept 2**	**Physician**
C0018834 (Heartburn)	C0018802 (Heart Failure, Congestive)	3.34
C0002962 (Angina Pectoris)	C0020621 (Hypokalemia)	3
C0010068 (Coronary Disease)	C0020621 (Hypokalemia)	3
C0018802 (Heart Failure, Congestive)	C0000737 (Abdominal pain)	3
C0002962 (Angina Pectoris)	C0000737 (Abdominal pain)	2.67
C0003811 (Arrhythmia)	C0018834 (Heartburn)	2.67
C0007192 (Cardiomyopathy, Alcoholic)	C0000737 (Abdominal pain)	2.67
C0007192 (Cardiomyopathy, Alcoholic)	C0020621 (Hypokalemia)	2.67
C0010068 (Coronary Disease)	C0000737 (Abdominal pain)	2.67
C0007192 (Cardiomyopathy, Alcoholic)	C0035238 (Respiratory System Abnormalities)	2.34
C0018799 (Heart Diseases)	C0000737 (Abdominal pain)	2.34
C0018799 (Heart Diseases)	C0035238 (Respiratory System Abnormalities)	2.34
C0018802 (Heart Failure, Congestive)	C0035238 (Respiratory System Abnormalities)	2.34
C0018834 (Heartburn)	C0035238 (Respiratory System Abnormalities)	2.34
C0018834 (Heartburn)	C0020621 (Hypokalemia)	2.34
C0000737 (Abdominal pain)	C0020621 (Hypokalemia)	2
C0003811 (Arrhythmia)	C0000737 (Abdominal pain)	1.67

Continued on next page

Concept 1	**Concept 2**	**Physician**
C0002962 (Angina Pectoris)	C0035238 (Respiratory System Abnormalities)	1.34
C0003811 (Arrhythmia)	C0035238 (Respiratory System Abnormalities)	1.34
C0010068 (Coronary Disease)	C0035238 (Respiratory System Abnormalities)	1.34
C0020621 (Hypokalemia)	C0035238 (Respiratory System Abnormalities)	1.34

Table C.2: Medical Concept Pairs provided by Caviedes and Cimino [Caviedes and Cimino, 2004].

References

Agarwal, P. and Searls, D. B. (2009). Can Literature Analysis Identify Innovation Drivers in Drug Discovery? *Nature Reviews. Drug Discovery*, 8(11):865–78.

Allan, J., Connell, M. E., Croft, W. B., Feng, F. F., Fisher, D., and Li., X. (2000). INQUERY and TREC-9. In *Proceedings of the 19th Text REtrieval Competition (TREC '00)*.

Allan, J., Croft, B., Moffat, A., and Sanderson, M. (2012). Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the 2nd Strategic Workshop on Information Retrieval in Lorne. *SIGIR Forum*, 46(1):2–32.

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic Models of Information Retrieval based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

Aronson, A. R. and Lang, F.-M. (2010). An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of American Medical Informatics Association*, 17(3):229–236.

Bai, J., Nie, J.-Y., Cao, G., and Bouchard, H. (2007). Using Query Contexts in Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, pages 15–22, New York, NY, USA. ACM.

Bai, J., Song, D., Bruza, P., Nie, J.-Y., and Cao, G. (2005). Query Expansion using Term Relationships in Language Models for Information Retrieval. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, pages 688–695, New York, NY, USA. ACM.

Balasubramanian, N., Kumaran, G., and Carvalho, V. R. (2010). Exploring Reductions for Long Web Queries. In *Proceedings of the 33rd International ACM SIGIR Conference on Research*

and Development in Information Retrieval (SIGIR'10), pages 571–578, New York, NY, USA. ACM.

Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36:673–721.

Battig, W. F. and Montague, W. E. (1969). Category Norms of Verbal Items in 56 Categories: A Replication and Extension of the Connecticut Category Norms. *Journal of Experimental Psychology Monographs*, 80(3, Pt. 2).

Bellman, R. E. (1961). *Adaptive Control Processes - A Guided Tour*. Princeton University Press, Princeton, New Jersey, U.S.A.

Bendersky, M. and Croft, W. B. (2009). Analysis of Long Queries in a Large Scale Search Log. In *Proceedings of the 2009 workshop on Web Search Click Data (WSCD'09)*, pages 8–14, New York, NY, USA. ACM.

Bendersky, M., Fisher, D., and Croft, W. B. (2011a). UMass at TREC 2010 Web Track: Term Dependence, Spam Filtering and Quality Bias. In *The 20th Text Retrieval Conference Proceedings (TREC'11)*. NIST. Special Publication.

Bendersky, M., Metzler, D., and Croft, W. B. (2011b). Parameterized Concept Weighting in Verbose Queries. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information (SIGIR'11)*, pages 605–614, New York, USA. ACM.

Berry, M. W., Drmac, Z., Elizabeth, and Jessup, R. (1999). Matrices, Vector Spaces, and Information Retrieval. *SIAM Review*, 41:335–362.

Beylkin, G. and Mohlenkamp, M. (2002). Numerical Operator Calculus in Higher Dimensions. *Proceedings of the National Academy of Sciences*, 99(16):10246–10251.

Beylkin, G. and Mohlenkamp, M. (2005). Algorithms for Numerical Analysis in High Dimensions. *SIAM Journal of Scientific Computing*, 26(6):2133–2159. Cited: 23.

Billerbeck, B. and Zobel, J. (2004). Questioning Query Expansion: An Examination of Behaviour and Parameters. In *Proceedings of the 15th Australasian Database Conference*

(ADC'04), volume 27, pages 69–76, Darlinghurst, Australia. Australian Computer Society, Inc.

Boystov, L. and Belova, A. (2011). Evaluating Learning-To-Rank Methods in the Web Track Adhoc Task. In *The 20th Text Retrieval Conference Proceedings (TREC'11)*. NIST. Special Publication.

Bruza, P. D. and Song, D. (2002). Inferring Query Models by Computing Information Flow. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02)*, pages 260–269, New York, NY, USA. ACM.

Buckley, C. (1995). Automatic Query Expansion Using SMART. In *Proceedings of the 3rd Text REtrieval Conference (TREC'95)*, pages 69–80.

Bullinaria, J. and Levy, J. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39:510–526.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming and SVD. *Behavior Research Methods*, 44:890–907.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes*, 25(2/3):211–257.

Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y. S., and Soffer, A. (2001). Static Index Pruning for Information Retrieval Systems. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, pages 43–50, New York, NY, USA. ACM.

Carpineto, C. and Romano, G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*, 44(1):1:1–1:50.

Caviedes, J. E. and Cimino, J. J. (2004). Towards the Development of a Conceptual Distance Metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85.

Charniak, E. (1994). *Statistical Language Learning*. MIT Press, Cambridge, MA, USA.

Cleverdon, C. (1970). *The Effect Of Variations In Relevance Assessments In Comparative Experimental Test Of Index Languages*. Cranfield Institute of Technology, UK.

Cleverdon, C., Mills, J., and M., K. (1966). *Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results*. College of Cranfield, UK.

Cohen, T., Schvaneveldt, R., and Widdows, D. (2010). Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. *Journal of Biomedical Informatics*, 43(2):240–256.

Cohen, T. and Widdows, D. (2009). Empirical Distributional Semantics: Methods and Biomedical Applications. *Journal of Biomedical Informatics*, 42(2):390–405.

Cohen, T., Widdows, D., Vine, L. D., Schvaneveldt, R. W., and Rindflesch, T. C. (2012). Many Paths Lead to Discovery: Analogical Retrieval of Cancer Therapies. In Busemeyer, J. R., Dubois, F., Lambert-Mogiliansky, A., and Melucci, M., editors, *The Quantum Interaction Conference (QI'12)*, volume 7620 of *Lecture Notes in Computer Science*, pages 90–101. Springer.

Collins-Thompson, K. (2009). Reducing the Risk of Query Expansion via Robust Constrained Optimization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, pages 837–846, New York, NY, USA. ACM.

Cormack, G. V., Smucker, M. D., and Clarke, C. L. (2011). Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. *Inf. Retr.*, 14(5):441–465.

Craswell, N. (2012). Evaluating Web Adhoc 2012 with Industry-Style Judging. In *The 21st Text Retrieval Conference Proceedings (TREC'12)*. NIST. Special Publication.

Croft, B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA.

Croft, W. B., Turtle, H. R., and Lewis, D. D. (1991). The Use of Phrases and Structured Queries in Information Retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'91)*, pages 32–45, New York, NY, USA. ACM.

Dennis, S. and Harrington, M. (2001). The Syntagmatic Paradigmatic Model: A Distributed Instance-based Model of Sentence Processing. In Isahara, M. and Ma, Q., editors, *Proceedings of the 2nd Workshop on Natural Language Processing and Neural Networks*, pages 38–45.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using Latent Semantic Analysis to Improve Access to Textual Information. *Proceedings of the SIGCHI conference on Human factors in computing systems*.

Dunn, M., Greenhill, S., Levinson, S., and Gray, R. (2011). Evolved Structure of Language shows Lineage-specific Trends in Word-order Universals. *Nature*, 473(7345):79–82.

Elffers, E. (2008). University of Amsterdam. *Du côté de chez Saussure*, page 79.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A Large-scale Model of the Functioning Brain. *Science*, 338:1202–1205.

Firth, J. R. (1957). *A Synopsis of Linguistic Theory 1930-55*, volume 1952-59. The Philological Society, Oxford.

French, R. M. and Labiouse, C. (2002a). Four Problems with Extracting Human Semantics from Large Text Corpora. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 316–322. Mahwah, NJ: Lawrence Erlbaum Associates.

French, R. M. and Labiouse, C. (2002b). Four Problems with Extracting Human Semantics from Large Text Corpora. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 316–322. Mahwah, NJ: Lawrence Erlbaum Associates.

Friedman, T., Casonato, R., and Logan, D. (2009). Innovation Forces that will change the Nature of Information Infrastructure. *Gartner Research*.

Glenberg, A. M. and Robertson, D. A. (2000). Symbol Grounding and Meaning: A Comparison of High-dimensional and Embodied Theories of Meaning. *Journal of Memory and Language*, 43:379–401.

Glenisson, P., Antal, P., Mathys, J., Moreau, Y., and Moor, B. D. (2003). Evaluation Of The Vector Space Representation In Text-Based Gene Clustering. In *Proceedings of the Pacific Symposium of Biocomputing*, pages 391–402.

Golub, G. and Reinsch, C. (1970). Singular Value Decomposition and Least Squares Solutions. *Numerische Mathematik*, 14:403–420.

Gorman, J. and Curran, J. R. (2006). Random Indexing using Statistical Weight Functions. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 457–464.

Grefenstette, G. (1992). Use of Syntactic Context to Produce Term Association Lists for Text Retrieval. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92)*, pages 89–97, New York, NY, USA. ACM.

Haas, W. (1973). John Lyons' 'Introduction to theoretical linguistics'. *Journal of Linguistics*, 9:71–113.

Harris, Z. (1954). Distributional Structure. *Word*, 10(23):146–162.

Harris, Z. (1968). *Mathematical Structures of Language*. Wiley, New York, USA.

Harter, S. P. (1975). A Probabilistic Approach to Automatic Keyword Indexing. Part I: On the Distribution of Specialty Words in a Technical Literature. *Journal of the American Society for Information Science*, 26(4):197–216. Cited: 168.

Hecht-Nielsen, R. (1994). Context Vectors: General Purpose Approximate Meaning Representations Self-organized from Raw Data. *Computational Intelligence: Imitating Life, IEEE Press*, pages 43–56.

Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. University of Twente.

Hoenkamp, E., Bruza, P., Song, D., and Huang, Q. (2010). An Effective Approach to Verbose Queries Using a Limited Dependencies Language Model. In *Advances in Information Retrieval Theory*, volume 5766 of *Lecture Notes in Computer Science*, pages 116–127. Springer Berlin / Heidelberg.

Holland, N. N. (1992). *The Critical I*. Columbia University Press, New York, USA.

Humphreys, M. S., Bain, J. D., and Pike, R. (1989). Different Ways to Cue a Coherent Memory System: A Theory for Episodic, Semantic, and Procedural Tasks. *Psychological Review*, 96:208–233.

Huston, S. and Croft, W. B. (2010). Evaluating Verbose Query Processing Techniques. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, pages 291–298, New York, NY, USA. ACM.

Ingwersen, P. (1992). *Information Retrieval Interaction*. Taylor Graham Publishing, London, UK, UK.

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Jacquemin, C. (1999). Syntagmatic and Paradigmatic Representations of Term Variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 341–348, Morristown, NJ, USA. Association for Computational Linguistics.

Jones, K. S. (2004). Idf Term Weighting and IR Research Lessons. *Journal of Documentation*, 60:521–523.

Jones, M. N. and Mewhort, D. J. K. (2007). Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, 114:1–37.

Kanerva, P., Kristoferson, J., and Holst, A. (2000). Random Indexing of Text Samples for Latent Semantic Analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036.

Karlgren, J. and Sahlgren, M. (2001). From Words to Understanding. In *In Uesaka, Y., Kanerva, P. and Asoh, H. (Eds.): Foundations of Real-World Intelligence*, pages 294–308. CSLI Publications.

Kolda, T. G. and Bader, B. W. (2009). Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., and Lawley, M. (2012). An Evaluation of Corpus-driven Measures of Medical Concept Similarity for Information Retrieval. In *The*

21st ACM International Conference on Information Knowledge Management (CIKM'12), page In Press.

Krovetz, R. (1993). Viewing Morphology as an Inference Process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 191–202, New York, NY, USA. ACM.

Lafferty, J. and Zhai, C. (2001). Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 111–119, New York, NY, USA. ACM.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104:211–240.

Lavrenko, V. (2004). *A Generative Theory of Relevance*. PhD thesis, University of Massachusetts Amherst. Director-Croft, W. Bruce and Director-Allan, James.

Lavrenko, V. and Croft, W. B. (2001). Relevance-Based Language Models. *In Proceedings of the 24th Annual ACM Conference of Research and Development in Information Retrieval (SIGIR'01)*, pages 120–127.

Lee, L. (1999). Measures of Distributional Similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99)*, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lin, J. and Gunopulos, D. (2003). Dimensionality Reduction by Random Projection and Latent Semantic Indexing. In *Proceedings of the Text Mining Workshop, at the 3rd SIAM International Conference on Data Mining*.

Littman, M., Dumais, S. T., and Landauer, T. K. (1998). Automatic Cross-Language Information Retrieval using Latent Semantic Indexing. In *Cross-Language Information Retrieval, chapter 5*, pages 51–62. Kluwer Academic Publishers.

Lowe, W. (2000). What is the Dimensionality of Human Semantic Space? In *Proceedings of the 6-th Neural Computation and Psychology Workshop*, pages 303–311. Springer Verlag.

Luhn, H. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2:159–165.

Lund, K. and Burgess, C. (1996). Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior research methods, instruments and computers*, 28:203–208.

Lv, Y. and Zhai, C. (2009a). A Comparative Study of Methods for Estimating Query Language Models with Pseudo Feedback. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM'09)*, pages 1895–1898, New York, NY, USA. ACM.

Lv, Y. and Zhai, C. (2009b). Adaptive Relevance Feedback in Information Retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management (CIKM'09)*, pages 255–264, New York, NY, USA. ACM.

Lv, Y. and Zhai, C. (2010). Positional Relevance Model for Pseudo-relevance Feedback. In *Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, pages 579–586, New York, NY, USA. ACM.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge University Press, London.

Machens, C. K. (2012). Nueroscience. Building the Human Brain. *Science*, 338(6111):1156–1157.

Maron, M. E. and Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM*, 7(3):216–244.

Metzler, D. (2007). *Beyond Bags of Words: Effectively Modeling Dependence and Features in Information Retreival*. Monograph, University of Massachusetts Amherst.

Metzler, D. and Croft, W. B. (2005). A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, pages 472–479, New York, NY, USA. ACM.

Metzler, D. and Croft, W. B. (2007). Latent Concept Expansion using Markov Random Fields. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval (SIGIR'07), pages 311–318, New York, NY, USA. ACM.

Metzler, D. and Zaragoza, H. (2009). Semi-parametric and Non-parametric Term Weighting for Information Retrieval. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval (ICTIR'09)*, pages 42–53, Berlin, Heidelberg. Springer-Verlag.

Meyer, C. D., editor (2000). *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Miller, D. R. H., Leek, T., and Schwartz, R. M. (1999). A Hidden Markov Model Information Retrieval System. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 214–221, New York, NY, USA. ACM.

Mitchell, J. and Lapata, M. (2008). Vector-based Models of Semantic Composition. *Proceedings of ACL-08: HLT*, pages 236–244.

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880):1191–1195.

Mizzaro, S. (1998). How Many Relevances in Information Retrieval? *Interacting With Computers*, 10:305–322.

Newman, S. (1952). Linguistics: Methods in Structural Linguistics. Zellig S. Harris. *American Anthropologist*, 54(3):404–405.

Ogilvie, P. and Callan, J. P. (2001). Experiments Using the Lemur Toolkit. In *Text REtrieval Conference*.

Ogilvie, P., Voorhees, E., and Callan, J. (2009). On the Number of Terms used in Automatic Query Expansion. *Information Retrieval*, 12(6):666–679.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. (1957). *The Measurement of Meaning*. University of Illinois Press.

Padó, S. and Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Pavel, T. C. (2001). *The Spell of Language: Poststructuralism and Speculation*. University of Chicago Press.

Pedersen, T., Pakhomov, S., Patwardhan, S., and Chute, C. (2007). Measures of Semantic Similarity and Relatedness in the Biomedical Domain. *Journal of Biomedical Informatics*, 40(3):288–299.

Perfetti, C. A. (1998). The Limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, 25:363–377.

Plate, T. A. (2001). Holographic Reduced Representations: Convolution Algebra for Compositional distributed Representations. *In Proceedings of the 12th International Joint Conference on Articial Intelligence.*

Ponte, J. M. and Croft, W. B. (1998). A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 275–281, New York, NY, USA. ACM.

Porter, M. F. (1997). *An Algorithm for Suffix Stripping*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Raden, N. (2005). Start making Sense: Get from Data to Semantic Integration. *Intelligent Enterprise*, 10.

Rapp, R. (2002). The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Rapp, R. (2003). Word Sense Discovery based on Sense Descriptor Dissimilarity. In *Proceedings of the 9th Machine Translation Summit*.

Rennie, J. D. M. and Jaakkola, T. (2005). Using Term Informativeness for Named Entity Detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR '05, pages 353–360, New York, NY, USA. ACM.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304.

Robertson, S. E. (1981). Term Frequency and Term Value. In *Proceedings of the 4th Annual International ACM SIGIR conference on Information storage and retrieval (SIGIR'81)*, pages 22–29, New York, NY, USA. ACM.

Robertson, S. E. and Sparck Jones, K. (1988). *Relevance Weighting of Search Terms*. Taylor Graham Publishing, London, UK, UK.

Robertson, S. E. and Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.

Rocchio, J. (1971). *Relevance Feedback in Information Retrieval*, pages 313–323. Prentice-Hall.

Rohde, D. L. T., Gonnerman, L. M., and Plaut, D. C. (2006). An Improved Model of Semantic Similarity based on Lexical Co-occurence. *Communications of the ACM*, 8:627–633.

Sahlgren, M. (2005). An Introduction to Random Indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*.

Sahlgren, M. (2006). *The Word Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces*. PhD thesis, Department of Linguistics, Stockholm University. Director-Croft, W. Bruce and Director-Allan, James.

Sahlgren, M., Holst, A., and Kanerva, P. (2008). Permutations as a Means to Encode Order in Word Space. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 23–26.

Sahlgren, M. and Karlgren, J. (2002). Vector-Based Semantic Analysis Using Random Indexing for Cross-Lingual Query Expansion. In *Revised Papers from the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems (CLEF'01)*, pages 169–176, London, UK, UK. Springer-Verlag.

Salton, G. (1968). Relevance Assessments and Retrieval System Evaluation. Technical report, Cornell University, Ithaca, NY, USA.

Salton, G. and Buckley, C. (1990). Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41:288–297.

Salton, G., Wong, A., and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.

Sandin, F., Emruli, B., and Sahlgren, M. (2011). Incremental Dimension Reduction of Tensors with Random Index. *Computing Research Repository (CoRR)*, abs/1103.3585.

Schütze, H. (1993). Word Space. In *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

Schütze, H. and Pedersen, J. (1993). A Vector Model for Syntagmatic and Paradigmatic Relatedness. In *Proceedings of the 9th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, Oxford, England.

Sebeok, T. (2001). *Signs: An Introduction to Semiotics*. University of Toronto Press.

Shadbolt, N., Hall, W., and Berners-Lee, T. (2005). The Semantic Web Revisited. *IEEE Intelligent Systems*, pages 895–902.

Shakery, A. and Zhai, C. (2013). Leveraging Comparable Corpora for Cross-lingual Information Retrieval in Resource-lean Language Pairs. *Information Retrieval*, 16:1–29.

Shannon, C., Petigara, N., and Seshasai, S. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423.

Sitbon, L., Bellot, P., and Blache, P. (2008). Evaluation of Lexical Resources and Semantic Networks on a Corpus of Mental Associations. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association.

Smalheiser, N. R. and Swanson, D. R. (1998). Using ARROWSMITH: a Computer-assisted Approach to Formulating and Assessing Scientific Hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149–153.

Smolensky, P. and Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammer*, volume 1. Cognitive Architecture. MIT Press.

Stegmann, J. and Grohmann, G. (2003). Hypothesis Generation Guided by Co-word Clustering. *Scientometrics*, pages 111–135.

Stone, B. P., Dennis, S. J., and Kwantes, P. J. (2008). A Systematic Comparison of Semantic Models on Human Similarity Rating Data: The Effectiveness of Subspacing. In *Proceedings of the 13th Conference of the Cognitive Science Society*. Cognitive Science Society.

Strohman, T., Metzler, D., Turtle, H., and Croft, W. B. (2005). Indri: a language-model based search engine for complex queries. Technical report, in Proceedings of the International Conference on Intelligent Analysis.

Symonds, M., Bruza, P., Sitbon, L., and Turner, I. (2011a). Modelling Word Meaning using Efficient Tensor Representations. In *Proceedings of the 25th Pacific Asia Conference on Language, Information, and Computation (PACLIC'11)*, pages 313–322.

Symonds, M., Bruza, P., Sitbon, L., and Turner, I. (2011b). Tensor Query Expansion: A Cognitive Based Relevance Model. In *Proceedings of the 16th Australasian Document and Computing Syposium (ADCS'11)*, pages 87–94. RMIT University(Melbourne).

Symonds, M., Bruza, P., Zuccon, G., Sitbon, L., and Turner, I. (2012a). Is the Unigram Relevance Model Term Independent?: Classifying Term Dependencies in Query Expansion. In *Proceedings of the 17th Australasian Document Computing Symposium (ADCS'12)*, pages 123–127, New York, NY, USA. ACM.

Symonds, M., Bruza, P. D., Sitbon, L., and Turner, I. (2012b). A Tensor Encoding Model for Semantic Processing. In *Proceedings of the 21st ACM International Conference on*

Information and Knowledge Management (CIKM'12), pages 2267–2270, New York, NY, USA. ACM.

Symonds, M., Zuccon, G., Koopman, B., and Bruza, P. (2013). QUT_Para at TREC 2012 Web Track: Word Associations for Retrieving Web Documents. In *The 21st Text Retrieval Conference Proceedings (TREC'12)*. NIST. Special Publication.

Symonds, M., Zuccon, G., Koopman, B., Bruza, P., and Nguyen, A. (2012c). Semantic Judgement of Medical Concepts: Combining Syntagmatic and Paradigmatic Information with the Tensor Encoding Model. In *Proceedings of the 10th Australasian Language Technology Workshop (ALTA'12)*, pages 15–22.

Tao, T. and Zhai, C. (2007). An Exploration of Proximity Measures in Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, pages 295–302. ACM.

Turney, P. D. (2008). A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, volume 1, pages 905–912, Stroudsburg, PA, USA. Association for Computational Linguistics.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Van Overschelde, J. P., Rawson, K. A., and Dunlosky, J. (2004). Category Norms: An Updated and Expanded Version of the Battig and Montague (1969) Norms. *Journal of Memory and Language*, 50(3).

van Rijsbergen, C. J. (1977). A Theoretical Basis for the use of Co-occurrence Data in Information Retrieval. *Journal of Documentation*, 33:106–119.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

van Rijsbergen, C. J. (2004). *The Geometry of Information Retrieval*. Cambridge University Press.

Voorhees, E. and Tong, R. (2011). Overview of the TREC Medical Records Track. In *Proceedings of the 20th Text REtrieval Competition (TREC'11)*, MD, USA.

Voorhees, E. M. (1994). Query Expansion using Lexical-semantic Relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Weeds, J. E. (2003). Measures and Applications of Lexical Distributional Similarity. Technical report, University of Sussex.

Widdows, D. (2004). *Geometry and Meaning*. Center for the Study of Language and Information/SRI.

Xu, J. and Croft, W. B. (1996). Query Expansion using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 4–11, New York, NY, USA. ACM.

Xu, Y., Jones, G. J., and Wang, B. (2009). Query Dependent Pseudo-relevance Feedback based on Wikipedia. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'09, pages 59–66, New York, NY, USA. ACM.

Yang, C. C., Yang, H., Tang, X., and Jiang, L. (2012). Identifying Implicit Relationships between Social Media Users to Support Social Commerce. In *Proceedings of the 14th Annual International Conference on Electronic Commerce (ICEC'12)*, pages 41–47, New York, NY, USA. ACM.

Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A Simple and Efficient Sampling Method for Estimating AP and nDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, pages 603–610, New York, NY, USA. ACM.

Yuret, D. (1998). *Discovery of Linguistic Relations using Lexical Attraction*. PhD thesis, Department of Computer Science and Electrical Engineering, MIT.

Zhai, C. and Lafferty, J. (2001). Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01)*, pages 403–410, New York, NY, USA. ACM.

Zhai, C. and Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.

Zuccon, G., Nguyen, A., Leelanupab, T., and Azzopardi, L. (2011). Indexing without Spam. In *Proceedings of the 16th Australasian Document and Computing Syposium (ADCS'11)*.

Index